Comparative and molecular characterisation of a schizophrenia susceptibility locus

Martin S Taylor

PhD
The University of Edinburgh
2001

Declaration

I declare that:

(a) This thesis has been composed by myself

(b) That the work is my own, except where otherwise stated

Martin S Taylor

October 2001

Acknowledgements

First of all I would like to thank my supervisors Rebecca Devon (the boss) and David Porteous (the big boss) for guidance, help and encouragement throughout the PhD. In particular David and Rebecca helped by giving me the freedom to develop and explore my ideas while directing me back from the tangential when necessary. The other boys in bioinformatics, Colin Semple and Stewart Morris have also had extremely significant roles in helping me develop the skills and ideas that recur throughout this thesis. For his help in converting me to a Perl programmer and letting me play with the system, Stewart more than deserves all of the pints I owe him. Kathy Evans has been like a third supervisor to me, thanks for all of your help, especially during the long thesis writing days.

Both the West Wing and subsequently the Medical Genetics Section have been stimulating and interesting environments to work, thanks to all who made it so. Particularly the t(1;11) group, Sheila Christie and Anne Doherty for tolerating me in the lab so well and "scary" Heather Davidson who turned out to be really nice after all.

Very little of the work presented in this thesis would have been possible without the biological and computational resources provided by the MRC funded UK-HGMP-RC. I am also indebted to the genome biology community for its "open source" attitude to the distribution of data, software and information; long may it continue.

A great deal of thanks also goes to my friends and family. Mum, Dad, Neil, Katie and my grandparents have provided a great deal of support for which I will always be indebted. The satisfaction of arguing science with Jim over a pint is reason enough to justify a career in science. Cheers to Jim, Kate and the rest of the Liverpool crowd. Thanks to Andie and Nic for suffering the thesis writing days with me and for all of the parties. A special thanks to Andie NPB for everything, time for the next adventure...

"Nothing in biology makes sense except in the light of evolution"

Theodosius Dobzhansky

Abstract

A substantial genetic contribution to the aetiology of schizophrenia and other major mental illnesses has been convincingly and repeatedly established by family, twin and adoption studies. However, phenotypic and genetic heterogeneity have severely hampered linkage and association studies, and consequently the molecular basis of the genetic contribution remains undefined. The use of cytogenetic abnormalities to identify disease loci is a well established technique that overcomes many of the problems of linkage and association studies. A balanced t(1;11)(q42;q14) translocation segregates in a large Scottish family (LOD = 7.1) with schizophrenia and related psychiatric disorders. At least three independent studies have also identified the 1q42 region of the genome as a susceptibility locus for major mental illness. The chromosome 1 breakpoint region now represents one of the best-supported loci for susceptibility to major mental illness. Two novel genes are directly disrupted by the chromosome 1 breakpoint, Disrupted-In-Schizophrenia 1 and 2 (*DISC1* and *DISC2*). The central hypothesis of this work is that genes directly disrupted by, or near to the chromosome 1 breakpoint contribute a significant susceptibility to major mental illness. This thesis set out to characterise *DISC1*, *DISC2* and neighbouring genes through comparative sequence analysis. Specifically, the research aimed to better define the locus, the genes, their functions and regulatory sequences, to evaluate the functional consequences of the translocation and how these may relate to the t(1;11) phenotype.

Human genomic sequence over the breakpoint region was assembled. The *DISC1* region of the *Fugu rubripes* genome was cloned and 45 kb of contiguous genomic sequence generated. The orthologous region of the mouse and chicken genomes was identified and characterised. A pipeline for preliminary genomic annotation and subsequent comparative genomic analysis was developed using the cystic fibrosis locus as a model, and subsequently applied to the *DISC1* locus. The method of "annotation anchored global sequence alignment" substantially increased the sensitivity in detection of biologically relevant conserved sequence motifs. Comparative genomic analysis, RT-PCR and cDNA clone identification were used to

construct a transcriptional map of the *Fugu* genomic region and refine the human transcription map. Conservation of synteny between 0.7 Mb of the human genome and 45 kb of the *Fugu* genome was demonstrated, with one boundary of synteny being clearly defined. The region of conserved synteny contained the genes Egg Laying Nine-1 (*EGLN1*), Translin Associated factor X (*TRAX*) and *DISC1* in both species.

EGLN1 was found to be a member of a previously undescribed gene family. The mouse and human members were identified and characterised. In addition, evolutionary evidence for a novel mechanism of host – pathogen interactions was discovered. TRAX and its homologue Translin were tentatively identified as members of a nucleic acid helicase family of proteins, providing a mechanistic basis for their known biological roles, and suggesting previously undescribed functional aspects of these proteins. *DISC1* was found to be rapidly evolving in both genomic structure and protein sequence, although three N-terminal motifs and blocks of coiled coil forming potential in the C-terminal half of the protein are conserved features, suggesting a general structure and function for the protein. Neither the antisense transcript *DISC2* nor the intergenic splicing of *TRAX* to *DISC1* are conserved in *Fugu*.

The work presented in this thesis has substantially enhanced understanding of the chromosome 1 breakpoint locus both at the genomic and encoded protein level. Two novel gene families have been defined and characterised, allowing a more complete evaluation of their functional candidacy in the aetiology of major mental illness. The sequence and clone resources resulting from this work also form the basis for protein functional studies and future characterisation of the locus in animal models.

Contents

Chapter 1

Introduction

Chapter 2

Materials and Methods

Chapter 3

Cloning the *DISC1* region from *Fugu rubripes*

Chapter 4

The *TRAX - DISC1* region in vertebrates

Chapter 5

Genomic sequence annotation

Chapter 6

Sequence analysis of *DISC1*

Chatper 7

DISC2 – the antisense transcript

Chapter 8

Sequence analysis of the *TRAX* gene

Chapter 9

The EGLN gene family

Chapter 10

Final Conclusions

List of tables

List of figures

Abbreviations

µCi	Microcuries
µg	Microgram
µl	Microlitre
µM	Micromolar
A	Adenosine
A. thaliana	*Arabidopsis thaliana*
aa	Amino acid
AAGSAL	Annotation anchored global sequence alignment
AceDB	A *C. elegans* database
At	*Arabidopsis thaliana*
ATP	Adenosine triphosphate
BAC	Bacterial artificial chromosome
BLAST	Basic local alignment search tool
C	Cytosone
CAT	Computed axial tomography
C. elegans	*Caenorhabditis elegans*
Ce	*Caenorhabditis elegans*
CFTR	Cystic fibrosis transmembrane conductance regulator
cM	Centimorgan
CTP	Cytosine triphosphate
D. melanogaster	*Drosophila melanogaster*
D. rerio	*Danio rerio*
DEPC	Diethyl pyrocarbonate
dH_2O	Distilled water
DISC1	Disrupted in Schizophrenia 1
DISC2	Disrupted in Schizophrenia 2
DMS	American diagnostic and statistical manual
Dr	*Danio rerio*
DTT	Dithiothreitol
E. coli	*Escherischia coli*
EDTA	Ethylenediamine tetra-acetic acid
EGLN	Egl-Nine
EIDE	Enhanced integrated drive electronics
EMBL	European Molecular Biology Laboratory
EST	Expressed sequence tag
FOLH2	Folate hydroase 2
F. rubripes	*Fugu rubripes*
Fr	*Fugu rubripes*
G	Guanine
Gb	Gigabase
GRM5	Metabotropic glutamate receptor 5
GSS	Genomic survey sequence
H. sapiens	*Homo sapiens*

HCl	Hydrochloric acid
HERV	Human endogenous retrovirus
HGMP-RC	Human Genome Mapping Project – Resource Centre
hnRNA	Heterogeneous nuclear ribonucleic acid
Hs	*Homo sapiens*
HSP	High scoring segment pair
HTG	High throughput genomic sequence
ICD	International Classification of Diseases
IHGSC	International human genome sequencing consortium
IPTG	Isopropylthio-β-D-galactoside
KAc	Potassium acetate
kb	Kilobase
kV	Kilovolts
l	Litre
L. esculentum	*Lycopersicon esculentum*
Le	*Lycopersicon esculentum*
LMP	Low melting point
LOD	Logarithm of odds
M	Molar
M. musculus	*Mus musculus*
mA	Milliamps
Mb	Megabase
MgCl2	Magnesium chloride
MLOD	Maximum logarithm of odds
Mm	*Mus musculus*
MRC	Medical Research Council
MRI	Magnetic resonance imaging
mRNA	Messenger ribonucleic acid
NaAc	Sodium acetate
NAAG	N-methyl-L-aspartyl-L-glutamate
$NaCl_2$	Sodium chloride
NaOH	Sodium hydroxide
ncRNA	Non-coding ribose nucleic acid
NMDA	N-methyl-D-aspartate
O. sativa	*Oryza sativa*
ORF	Open reading frame
Os	*Oryza sativa*
P. aeruginosa	*Pseudomonas aeruginosa*
PAC	P1 artificial chromosome
PCNXL2	Pecanex like - 2
PCP	Phencyclidine
PCR	Polymerase chain reaction
PET	Positron emission tomography
RACE	Rapid amplification of cDNA ends
RAM	Random access memory
R. norvegicus	*Rattus norvegicus*
Rn	*Rattus norvegicus*
S. cerevisiae	*Saccharomyces cerevisiae*

S. pombe	*Schizosaccharomyces pombe*
Sc	*Sacharomyces cerevisiae*
Sec	Second
SNP	Single nucleotide polymorphism
Sp	*Schizosaccharomyces pombe*
SRS	Sequence retrieval system
T	Thymine
T. aestivum	*Triticum aestivum*
T. nigroviridis	*Tetraodon nigroviridis*
Ta	*Triticum aestivum*
Tn	*Tetraodon nigroviridis*
TRAX	Translin associated factor X
UTR	Untranslated region
V	Volts
V. cholerae	*Vibrio cholerae*
Vc	*Vibrio cholerae*
X. laevis	*Xenopus laevis*
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactoside
Xl	*Xenopus laevis*
YAC	Yeast artificial chromsome

Chapter 1

Introduction

1.1 Preface

Schizophrenia is a severe and debilitating mental illness that is among the most common causes of chronic morbidity world wide (Lopez and Murray, 1998). Its symptoms include delusions and hallucinations that are considered to be caused by an excess of normal brain function, social withdrawal, impaired volition and speech that are considered to be due to a deficit of normal brain function. Schizophrenia was first defined as clinical entity during the 18th century by Emil Kraepelin, although it is speculated that schizophrenia has been evident in the human population for thousands of years (Jeste *et al.*, 1985). Initially termed dementia praecox, Bleuler (1950) coined the name schizophrenia (from Greek: *skhizo* "to split" and *phren* "mind") to signify the loss of integration between emotions, thought and actions; not as is often miss conceptualised the splitting of one personality into two. The schizophrenia phenotype is highly variable between individuals and even in the early days of psychiatric research it was recognized that schizophrenia is not a monadic condition, but is likely to represent several pathological entities.

Substantial personal, social and economic burdens are imposed by schizophrenia. In the UK alone there are 185,400 people receiving medical treatment for schizophrenia at an average cost of £2,138 per year per patient (1994 figures; Hall, 1997). Direct and indirect costs of schizophrenia in the UK are estimated at more than £1.7 billion per year (Hall, 1997). Within the US, the financial cost of schizophrenia was estimated to be $65 billion when direct treatment, societal and family costs are taken into account (Wyatt and Glazer, 1996). The emotional costs of schizophrenia are undoubtedly substantial, but are difficult to quantify. Suicide and attempted suicide are significantly more common in schizophrenics than in the general population. Ten

percent of clinically diagnosed schizophrenics are estimated to commit suicide, and attempted suicide rates are estimated to lie between 20 and 50% (Siris, 2001).

There have been many recent advances in the treatment of schizophrenia and a range of neuroleptics are now available to ameliorate the symptoms of disease. However, as the underlying causes of the illness are unknown it is only possible to treat the symptoms, not the cause. It is also the case that up to 30% of suffers are unresponsive to available treatments. As the response to a particular neuroleptic cannot be reliably predicted from the presentation of symptoms, the therapeutic strategy for the application of neuroleptics is one of trial and error. There are severe side effects associated with several of the neuroleptics, including tardive dyskinesia (uncontrollable movements) that may persist even after withdrawal of medication. Such side effects are of great concern particularly as amelioration of the psychiatric symptoms cannot be guaranteed.

It is probable that schizophrenia represents several aetiologically distinct diseases. Observations of abnormal brain chemistry, brain structure and neural patterning indicate that the symptoms of schizophrenia may be manifestations of underlying structural and functional disruption (section 1.2.3). Family, twin and adoption studies have repeatedly and convincingly demonstrated that there is a substantial genetic contribution to schizophrenia susceptibility (section 1.3.4). The combination of evidence for both abnormalities in the schizophrenic brain and a genetic contribution to the aetiology of schizophrenia argues that there is likely to be a molecular basis for disease susceptibility. Identification of the molecular risk factors will provide therapeutic targets for rational drug design. In addition such factors may allow aetiologically relevant categorisation of mental illnesses and be of use for pharmacogenomic (Adam *et al.*, 2000) therapeutic strategies.

It is likely that genetic and phenotypic heterogeneity as well as epistatic and gene – environment interactions have severely hampered attempts to define the molecular basis of schizophrenia and major mental illness as a whole. A large Scottish family has been reported that segregates schizophrenia and related mental illnesses through

five generations (StClair *et al.,* 1990; Blackwood *et al.,* 2001). Using a single large family for linkage studies minimises the problems associated with aetiological heterogeneity. A balanced chromosomal translocation was also found to segregate through the large Scottish family. Significant (MLOD = 7.1) co-segregation of the chromosomal aberration and mental illness (StClair *et al.,* 1990; Blackwood *et al.,* 2001) provides a rare opportunity to define the molecular basis of schizophrenia susceptibility. The work presented in this thesis builds on these findings.

1.2 The schizophrenia phenotype

Schizophrenia is a clinical syndrome encompassing several discrete clinical features. No single symptom is unique to schizophrenia or manifest in every case. The core symptoms are defined as positive (psychotic) or negative, reflecting a perceived excess or deficiency of normal brain function. Positive symptoms include hallucinations, delusions, disorganised thought, alteration of social functioning and loosening of association. The negative symptoms include neuropsychological deficits (section 1.2.2) loss of emotional responsiveness, social withdrawal, impaired speech and impaired volition. A diagnosis of schizophrenia can be based on positive, negative or both types of symptoms (box 1.1), although in most cases cognitive impairment (considered a negative symptom) precedes psychosis (Rund and Borg, 1999).

1.2.1 Diagnosis

One of the major confounding factors in genetic investigations of schizophrenia, is the heterogeneity of clinical manifestation. This same problem is encountered in defining standardised diagnostic criteria. With no core symptom or adequate biological marker, clinical diagnosis is based on psychiatric evaluation at clinical presentation. A psychiatric evaluation is usually in the form of a structured interview between the patient and a psychiatrist. Diagnostic criteria have been developed in an attempt to standardise psychiatric diagnosis, providing a common language within psychiatry and allowing informative comparison between studies.

Currently the most commonly used diagnostic criteria are Diagnostic and Statistical Manual of Mental Disorders (DSM) (American Psychiatric Association, 1994),

Research Diagnostic Criteria (RDC) (Spitzer 1978) and Classification of Mental and Behavioral Disorders (ICD) (World Health Organization, 1992). DSM is currently on its 4th major revision and ICD on its 10th. While still subjective, the interrater reliability (consistency of diagnosis on independent assessment) of both ICD-10 and DSM-IV have been thoroughly tested in large scale studies and were found to be robust (Hiller *et al.,* 1994). However, the aetiological validity of these classifications is more questionable (Strober *et al.,* 1995).

A set of characteristic symptoms have been defined of which typically two or more must be met in most criteria for a diagnosis of schizophrenia. Box 1.1 outlines the formal DSM-IV diagnostic criteria for a psychiatric diagnosis of schizophrenia.

DSM-IV Diagnostic criteria for schizophrenia

(a) Characteristic symptoms

Two or more of the following. Only one symptom is required if delusions are bizarre or hallucinations consist of a voice keeping up a running commentary on the persons behaviour or thoughts, or two or more voices conversing with each other.

(1) Delusions
(2) Hallucinations
(3) Disorganised speech
(4) Grossly disorganised or catatonic behaviour
(5) Negative symptoms

(b) Social / occupational dysfunction

For a significant portion of the time since the onset of the disturbance, one or more major areas of functioning such as work, interpersonal relations, or self-care are markedly below the level achieved prior to the onset (or when the onset is in childhood or adolescence, failure to achieve expected level of interpersonal, academic, or occupational achievement).

(c) Duration

Continuous signs of the characteristic symptoms for at least six months, possibly including periods of the symptoms present in an attenuated form.

(d) Schizoaffective and mood disorder exclusion

An absence of major depressive, manic or mixed episodes occurring concurrently with the characteristic symptoms or if mood episodes have occurred concurrently their total duration has been brief relative to the characteristic symptoms.

(e) Substance / general medical condition exclusion

The disturbance is not due to the direct physiological effects of a substance or a general medical condition.

(f) Relationship to a pervasive developmental disorder

If there is a history of autistic disorder or another pervasive developmental disorder, the additional diagnosis of schizophrenia is made only if prominent delusions or hallucinations are also present for at least one month

Box 1.1; DSM-IV diagnostic criteria. Adapted from DSM-IV (American Psychiatric Association, 1994).

Beyond the primary diagnosis of schizophrenia, attempts are made to categorise the diagnosis into a subtype. The subtype categories used by both and DSM-IV are based on the original subdivisions defined by Kraeplin (1919). Several additional subdivisions have subsequently been added to both ICD and DSM. The DSM-IV criteria recognises five sub-types: paranoid, catatonic, disorganised, undifferentiated and residual. The evidence for these subtypes to breed true is weak and the observation that a patient can change subtypes over the course of illness (Kendell, 1987) suggests that the subtypes are reflective of environmental rather than genetic aetiological factors. Attempts have been made at constructing more genetically relevant subdivisions of schizophrenia (Leonhard, 1979; Crow, 1980). The system of Leonhard having gained support from family studies of Strober *et al.,* (1995).

While it is apparent that not all clinical presentations of schizophrenia reflect a single genetic aetiology, it is equally the case that all clinical presentations of a common genetic factor are not diagnosed as schizophrenia. The schizophrenia spectrum constitutes a range of related psychiatric diagnoses that, while not conforming to the strict diagnostic criteria for schizophrenia, show clear familial clustering and heritability with schizophrenia (section 1.3). Originating as a concept from an adoption study by Kety *et al.,* (1988), the schizophrenia spectrum was first referred to as latent or borderline schizophrenia. Psychiatric diagnoses proposed to be included in the spectrum include schizoaffective disorder, atypical psychosis, delusional disorder and personality disorders (schizotypical, schizoid and paranoid) (Prescott & Gottesman, 1993). Given the probable aetiological heterogeneity of schizophrenia (sections 1.3.5), attempts at a rigid definition of the schizophrenia spectrum appears doomed to failure, as distinct causative factors may manifest as overlapping spectra of symptoms and diagnosis. It is for these reasons that the disease spectrum is often defined for specific genetic studies, particularly if parametric statistical methods are being used. The consequent problems of multiple testing and statistical validation are discussed in section 1.5.3.

1.2.2 Neuropsychological phenotype

The most consistently replicated neuropsychological deficits in schizophrenia patients have been observed in attention, information processing, learning or recall, and executive functions such as planning and decision making (Michie, 2000; Green 2000). The neuropsychological deficits of schizophrenia are considered negative symptoms of the schizophrenia phenotype.

Attention and information processing deficits relate to the ability to focus attention on a specific stimulus, divide processing between multiple stimuli, particularly if these are received through separate channels (e.g. a visual and an auditory stimulus) and the ability to sustain processing over time. A frequently used measure of such processing abilities is the span of apprehension (SOA) test, where subjects must determine whether for example a "T" or an "F" letter is amongst a group of otherwise random letters shown briefly to the subject. A consistent finding of this type of test is that schizophrenics and well matched controls perform equally well with small numbers of "distracter" letters, but with increasing array sizes there is increasing polarization between the two groups, with schizophrenics accurately reporting fewer targets than controls (Bryson *et al.,* 2001; Chang and Lenzenweger, 2001). These findings are well supported by independent experimental designs (Rund and Borg, 1999 for review).

A consensus of these investigations into attention and information processing, strongly suggest that schizophrenics perform as well as normal controls when the complexity of the task is low, but as complexity increases with added "noise" schizophrenics progressively fail to perform as well as the controls. These observations have lead to the hypothesis of schizophrenics being relatively unable to focus attention by filtering extraneous information, and consequently overloading their processing resources (Asarnow and Granholm, 1991; Jeste *et al.,* 1996). A further, and consistently reported observation is that these deficits are also found in schizophrenic subjects in a remitted state, biological relatives of schizophrenic patients and individuals with premorbid markers for schizophrenia spectrum mental

illness (Asarnow and Granholm, 1991) leading to the intriguing possibility of defining core deficits of this phenotypically heterogeneous condition.

Learning and memory have been investigated, again comparing sets of schizophrenics to controls. Several independent studies have found consistent impairment of schizophrenics on recall and recognition tasks, usually utilizing word list based testing procedures (Silva *et al.,* 2000; Lussier and Stip, 2001). These deficits are particularly pronounced when semantic or other associative operations are required to enhance performance. Although there is a clear deficiency in the ability to learn verbal and declarative information, there is no evidence for such learning problems in procedural skills involving motor and problem-solving (Jeste *et al.,* 1996).

Executive functions are those neural systems involved in reacting and adapting to the environment, such as the preparation, initiation and regulation of action, abstract reasoning and hypothesis testing. Impaired performance of schizophrenics relative to controls are reported in several standardised tests that probe executive functions. Such tests include the Winsconsin card sorting task and Benton's controlled oral word association tasks, as well as tests of impaired memory for temporal ordering of information (Jeste *et al.,* 1996 for review).

Even with this wealth of studies into neurophsychological deficits associated with schizophrenia, it is observed that up to 60% of schizophrenic patients demonstrate intact performance in these tests (Jeste *et al.,* 1996 for review), again pointing towards the heterogeneous nature of this condition. The relevance of the detected phenotypes may lie in the ability to subgroup schizophrenics. The power of this approach has already been demonstrated by the ability to predict to some extent the response of patients to neuroleptics (Smith *et al.,* 1992; Robinson *et al.,* 1999). The apparent familial clustering of phenotypes and presentation even in premorbid and remitted states could add considerable power to family studies, such as this project through defining a family specific core phenotype to be scored against genotype, as a means of reducing phenocopy noise.

1.2.3 Neuropathology

A wide range of brain imaging technologies have been used to investigate the possible association of structural brain abnormalities or disruption of normal brain function with schizophrenia. Computed tomography (CT) and magnetic resonance imaging (MRI) have been used extensively to investigate brain structure. Positron emission tomography (PET) and a host of other fundamentally similar techniques have been used to investigate blood flow in the brain in response to stimuli, a correlate of regional changes in neural activity. More recently, PET and MRI technologies have been combined to produce merged functional and structural information with good temporal and spatial resolution.

Structural differences between schizophrenic and non-schizophrenic brains were first reported by Johnstone *et al.,* (1976) using computerised axial tomography and have since been replicated many times in standard and more sophisticated case control studies (Okazaki, 1998 for review). In a review of the greater than 40 studies that have found significant structural differences between schizophrenics and non-schizophrenics Lawrie and Abukmeil (1998) have calculated median percentage volume differences in the implicated brain regions (table 1.1).

Brain region	Left	Right
Whole brain	97%	
Temporal lobe	94%	90.5%
Amygdala/hippocampal complex	93.5%	94.5%
Parahippocampus	86%	91%
Lateral ventricles	144%	136%

Table 1.1; Percentage of brain region volume in schizophrenic relative to normal brain. Adapted from Lawrie and Abukmeil (1998).

It is a consistent finding that structural brain abnormalities associated with schizophrenia are more pronounced in males and on the left side of the brain. A

further consistency is the reduction in volume of brain tissue, this is especially pronounced by the increase in ventricle volume, the most consistently reported structural abnormality. The ventricles are fluid filled spaces in the brain, enlargement of which suggests concomitant reduction of brain volume. Segmentation studies suggest that reductions in volume reflect reductions in grey matter and that white matter might actually be increased in schizophrenic relative to non-schizophrenic brains (Wright *et al.,* 1999).

Increased ventricular to brain volume ratio (VBR), appears to be evident prior to the onset of clinical psychosis and is non-progressive in course (Jaskiw *et al.,* 1994; Harrison, 1999 for review), hinting at a neurodevelopmental rather than neurodegradative aetiology. A neuroimaging study of monozygoytic twins discordant for schizophrenia found increased VBR in schizophrenics and their twin, but not in their discordant sibs (Suddath *et al.,* 1990). Similar results were also observed in a more rigorously controlled, but smaller study (Ohara *et al.,* 1998). The variance in VBR values between monozygotic twins concordant and those discordant for schizophrenia were not statistically significant (Ohara *et al.,* 1998). VBR variances between non-schizophrenic monozygotic twins were significantly reduced relative to those with at least one presentation of schizophrenia (Ohara *et al.,* 1998). These results were considered to demonstrate that the underlying neuropathological processes cannot be solely genetic (Sudath *et al.,* 1990; Ohara *et al.,* 1998). Although it is clear that the aetiology of schizophrenia is not solely genetic (section 1.3.4), the emphasis of these conclusions appear to have been misdirected. The results of Sudath *et al.,* (1990) and particularly Ohara *et al.,* (1998) suggest that although VBR values themselves may be particularly prone to environmental influences, the degree of plasticity appears to be highly heritable and directly related to the incidence of schizophrenia in twin pairs.

As well as the gross volumetric differences in schizophrenic compared to control brains, a variety of cytoarchitectural abnormalities in have been observed (Harrison, 1999 for review). These include anomalies of neuronal positioning (Akbarian *et al.,* 1996; Falkai *et al.,* 2000; McGalshan and Hoffman, 2001 for review), neuron size

(Arnold *et al.,* 1995) and density (Benes *et al.,* 1998). The cytoarchitectural abnormalities and the absence of gliosis (Harrison, 1999 for review) associated with schizophrenia have lead to the prevailing theory of schizophrenia as a neurodevelopmental rather than neurodegenerative pathology (Harrison 1999; Lobato *et al.,* 2001).

Functional brain imaging investigations have frequently reported abnormalities of cerebral perfusion and glucose metabolism in patients with schizophrenia. As with all of the previously discussed phenotypes of schizophrenia, these abnormalities are manifest in only a subset of patients. However, in this case there appears to be a correlation between the nature of the clinical presentation and neuropathological manifestation. Schroder *et al.,* (1995) demonstrated a relationship between frontal lobe hypometabolism (hypofrontality) and the negative psychiatric symptoms of schizophrenia (section 1.2.1), an observation that had previously been suggested by several groups (Jeste *et al.,* 1996 for review).

Hypofrontality is by far the most thoroughly investigated functional neuropathology of schizophrenia and has been supported in neuroleptic naive (Buchsbaum *et al.,* 1992), neuroleptic free and medicated patients (Farkas *et al.,* 1984). Hypofrontality is either controlled for in a case control manner in the resting condition, or more rigorously using a relative measure of defining a baseline with control tasks and measuring frontal lobe activity with prefrontal tasks. Warkentin *et al.,* 1990 found relative hypofrontality with schizophrenic patients in remission.

1.2.4 Biological markers for schizophrenia

Problems with the clinical heterogeneity of schizophrenia (section 1.2) and indications of latent or schizophrenia spectrum diagnosis (section 1.2) have prompted the search for quantifiable biological markers of schizophrenia susceptibility. Such markers could be of direct clinical relevance through the aetiologically appropriate sub-typing of patients. In family studies, a reliable biological marker could be used to quantifiably score individuals in a manner independent of psychiatric diagnosis. For clinical use, a biometric measure is generally required to distinguish affected from unaffected individuals by at least three standard deviations. To be of benefit for

genetics studies, two standard deviations would be adequate. A useful biological marker for schizophrenia would be expected to posses the following properties: (1) heritability; (2) association with disease at the population level; (3) presentation even when clinical symptoms are in remission; (4) segregation of the illness in families of ill probands.

Smooth pursuit eye tracking

An object moving smoothly across a visual field is followed by movements of the eye to maintain an image of the object at the centre of the retina. This eye tracking is smooth and lags slightly behind the movement of the object. In other circumstances, eye movements can be rapid and abrupt (saccadic) for example, when visual attention is refocused on a new object. A disruption of the normal smooth pursuit has been observed in 51 to 81% of schizophrenics compared to 8% of non-schizophrenic controls (Holzman *et al.,* 1977; Blackwood *et al.,* 1991). The disruption found in schizophrenic subjects consists of saccadic intrusion when smooth pursuit would be the norm. The results of family and twin studies have shown the eye tracking defect to be highly heritable (Blackwood *et al.,* 1991; Katsnis *et al.,* 2000) and it has been proposed as a manifestation of the same underlying pathology as schizophrenia (Matthysse *et al.,* 1986).

Event related potentials

Externally measured (scalp attached electrodes) electrical potentials within the brain have been widely reported as abnormal in schizophrenia (Sengoku and Takagi, 1998 for review). While brain electrical potentials are generally problematic to measure, interpret and compare, the event related potentials represent an experimentally tractable sub-set of these signals. Event related potentials (ERPs) are a series of brain electrical signals induced by a specific stimulus (auditory, visual or somatosensory). Those signals that occur very rapidly following the stimulus (0 – 50 ms) are thought to represent low level brain function. In contrast, those signals with a greater latency are thought to reflect higher executive brain functions (D. Blackwood, personal communication).

In schizophrenics compared to controls, the most consistent ERP differences are in a decrease in amplitude of signals and an increased time between stimulus and signal peaks. The most pronounced differences have been observed with the P300 ERP that typically occurs approximately 300 ms following auditory or visual stimulus (Ford, 1999 for review). The P200 ERP has also been shown to have both decreased amplitude and increased latency in schizophrenics compared to controls (McCarley *et al.,* 1991). These patterns of P200 and P300 ERP have been found to be unaffected by neuroleptic medication and are independent of clinical state at the time of testing (Blackwood *et al.,* 1987; Mathalon *et al.,* 2000). These disruptions of P300 ERP are highly heritable and co-segregate with schizophrenia and other major mental illnesses in familes (Lenox *et al.,* 1945; Vogel, 1970; Blackwood *et al.,* 1991; Blackwood *et al.,* 2001; Bharath *et al.,* 2000 for review). The significance of P300 ERP is discussed further in section 1.6 as this is a marker that segregates exceptionally well with both mental illness and karyotype in the family introduced in section 1.6.

Prepulse inhibition

Sensorimotor gating is the neurological mechanism of filtering the milieu of stimuli (visual, olfactory, auditory and somatosensory) and being able to focus attention on a sub-set of the information. In 1960, Venables proposed that this mechanism was disrupted in schizophrenia causing a "flooding" by sensory inundation. A similar general mechanism has also been implicated in Huntington's disease (Swerdlow *et al.,* 1995; Tourette's syndrome (Castellanos *et al.,* 1996) and obsessive compulsive disorder (Swerdlow *et al.,* 1993), collectively considered "gating disorders".

Prepulse inhibition (PPI) of the startle response is considered an operational measure of sensorimotor gating. If an intense acoustic stimulus (startle pulse) is applied without warning to an individual they "jump", this is the startle response. If a low intensity, non-startling acoustic stimulus is applied shortly prior to the startle pulse, the extent of the startle response is substantially attenuated. The timing of the prepulse is crucial in the elicitation of PPI, being optimal 60 to 120 ms prior to the startle pulse (Linn and Javitt, 2000). In humans with gating disorders including

schizophrenia this prepulse inhibition is deficient (Swerdlow *et al.,* 1999; Geyer *et al.,* 2001 for review).

As for smooth pursuit of eye tracking and P300 ERP, the phenotype of PPI is not restricted solely to schizophrenia, is not evident in all schizophrenia patients and is not sufficiently discriminatory to be of use in clinical diagnosis (Thaker, 2000). However, there is particular interest in PPI as it is deficient in pharmacologically induced human models of schizophrenia (section 1.4). PPI deficiency is also observed in animals treated with pharmacological agents that induce schizophrenia like symptoms and PPI deficiency in humans. These findings suggest that PPI could provide a means of testing proposed animal models of mental illness (section 1.4 for further discussion).

1.3 A genetic basis

It is now widely accepted that there is a substantial genetic component to schizophrenia susceptibility, demonstrated through family, twin and adoption studies. It is also clear that the disease does not generally segregate in a monogenic Mendelian manner. Clues as to the true phenotypic spectrum of schizophrenia have also been determined from studies into its genetic basis.

Under strict diagnostic criteria, the lifetime morbid risk of schizophrenia is at a relatively constant 1%, independent of geographical location or race (Jablensky *et al.,* 1992). This epidemiological distribution is atypical for genetic traits, which are often subject to founder effects with consequent racial and geographical clustering such as is observed, for example, sickle cell anaemia, Tay Sachs and cystic fibrosis (OMIM: 603903; 272800; 219700 respectively). There is some evidence for geographical hot spots of schizophrenia in Sweden, Yugoslavia (prior to political subdivision) and Ireland (Wyatt *et al.,* 1988, Eaton 1991). However, these results may have been biased by differences in diagnostic practices.

1.3.1 Family studies

If there is a substantial genetic component to the aetiology of schizophrenia, familial clustering would be expected with risk proportional to relatedness of subject to

proband. The vast majority of family studies have demonstrated this relationship. The few that have failed to find evidence of familial clustering (Pope *et al.,* 1982 and Abrams & Taylor, 1983) have been heavily criticised on methodological grounds (Moldin, 1997). In a classical family study, Waddington and Youssef (1996) reported a 6.1% risk of schizophrenia in first degree relatives of probands and risk amongst sibs was 8.3%, well above that of the general population at 1%. Inconsistency of diagnostic criteria and methodology are expected to be the primary cause of variation in reported figures of risk to relatives and these factors also complicate attempts to make direct comparisons between studies. Despite these caveats, the majority of family studies have found patterns of risk similar to those described by Waddington and Youssef (McGue & Gottesman, 1991; and Moldin, 1997 for review).

1.3.2 Twin studies

The familial clustering of schizophrenia and more prominently schizophrenia spectrum disease implies a genetic risk factor. Further support for a genetic component to schizophrenia aetiology is derived from twin studies. Monozygotic (MZ) twins are genetically identical; where as dizygotic (DZ) twins are expected to be 50% identical by decent. It is also normal for twins to share a common intrauterine environment. Twin studies are consequently very powerful methods of investigating the extent of genetic contribution to a phenotype. Twin studies are assessed in terms of concordance rates, the probability that a twin is affected given that the co-twin is affected.

Twin concordance studies have been carried out since 1920 (Kringlen, 2000 for review). Several large scale studies made use of well documented national twin and health records from Scandinavian countries (Kringlen *et al.,* 1966; Fischer *et al.,* 1969; Tienari *et al.,* 1994). Tienari reported MZ concordance at 15%, DZ concordance at 7%; Kringlen found an MZ concordance of 31% and DZ of 8%; MZ concordance of 24% and DZ at 10% were observed by Fischer. In a large scale meta analysis, covering all methodologically sound twin studies reported between 1920 and 1987, Moldin (1998) found mean concordance rates for MZ twins to be 46% and only 14% for DZ twins. Consistent with the results of Moldin, Prescott *et al.,* (1993) proposed a median concordance rate for MZ twins of 48% and 15% for DZ twins,

based on several studies with like methodology. It is likely that the varying diagnostic criteria are principally to blame for the differences in concordance between studies.

A little commented upon consistency between studies is the difference between MZ and DZ concordance rates, DZ concordance rates are generally reported at less than 50% of the concordance rate of MZ twins. Dizygotic twins are expected to be genetically 50% identical by decent and environmental differences between twins are likely to be minimal irrespective of zygosity, the observed differences between MZ and DZ concordance rates can be considered suggestive of a polygenic aetiology. Fischer, (1971) presented data that suggested the offspring of an unaffected twin of a discordant monozygotic pair had the same risk of schizophrenia as the offspring of the affected twin. The findings of Fischer complement classical twin studies well, and are entirely consistent with a genetic aetiology and partial penetrance. The partial penetrance provides strong evidence for environmental as well as genetic contributions to the aetiology of mental illness.

1.3.3 Adoption studies

Family and twin studies provide strong evidence for a genetic susceptibility to schizophrenia. However, in the family and twin stuides discussed, there were only limited means to control for sociological and other environmental influences. A variety of adoption study strategies have been used to address these issues. "Adoptees studies" were used to determine the prevalence of schizophrenia in adoptees born to schizophrenic parents compared with those born to non-schizophrenic parents. In the first study of this type, Heston, (1966) found schizophrenia diagnosed in 5 of 47 adoptees born to schizophrenic mothers compared to none in the control group of 50 individuals.

The "adoptee relatives study" design was used by Kety *et al.,* (1971; 1976) and Tienari *et al.,* (1994). In this study design, prevalence of schizophrenia is measured in the biological and adopted relatives of a schizophrenic proband adoptee. Tienari *et al.,* (1994) reported an excess of schizophrenia in biological, but not adopted families. Although consistent with other findings, this study would have been more

robust with a control set of non-schizophrenic adoptees. The proposed control set would control for altered rates of schizophrenic parents offering or accepting children for adoption.

Rosenthol *et al.,* (1978) reported an adoptee relatives study, modified to investigate rates of schizophrenia in paternal half sibs of schizophrenic probands. Paternal half sibs that have been reared in distinct environments will be genetically 25% identical by decent, but will not have shared prenatal, perinatal or neonatal environments. The finding of excess (13.8%) schizophrenia and related disorders in paternal half sibs compared to paternal half sibs of controls in this study, again strongly implies a biological heritability of schizophrenia.

1.3.4 A substantial genetic contribution to schizophrenia susceptibility

While no one study design alone is sufficient to demonstrate unambiguously a genetic aetiology, the various designs of family, twin and adoption studies complement each other. The gestalt of these studies is compelling evidence for a major genetic contribution to schizophrenia susceptibility. The genetic epidemiology of schizophrenia was investigated in 2,495 MZ and 5,378 same-sex DZ twin pairs (All same-sex Finnish twins born between 1940 and 1957) using structural equation modeling (Cannon *et al.,* 1998). The outcome of Cannon's investigation suggested 83% of variance in liability is due to additive genetic factors, consistent with findings of Cardno *et al.,* (1999) who reported a schizophrenia heritability estimate of 85% using index twin pairs and standardized diagnostic criteria.

1.3.5 Genetic heterogeneity

It is apparent from both family and twin studies that the inheritance of schizophrenia cannot be explained by a single genetic locus for the majority of families segregating the disease. Risch and Baron (1984) proposed that a polygenic or mixed model is most consistent with family data, lifetime disease incidence and twin concordance rates. Risch (1990) proposed an estimate of 2 or 3 epistatic loci based on a meta analysis of Western European studies. Murray *et al.,* (1985) carried out a meta analysis of family and twin studies that suggested considerable genetic heterogeneity; strongly arguing that large families will be the most useful for

molecular genetic investigation. This view has subsequently been reiterated in a number of publications (Franzek and Beckmann, (1998) and Garver (1997) for review). The argument for genetic heterogeneity would seem a sound assumption given the phenotypic spectrum of the disease (section 1.2), variation in neuropatholgoy between families (section 1.2.3) and apparent differencies in the mode of inheritance between families (Garver, 1997).

1.4 Models of schizophrenia

Several pharmacological and animal models of schizophrenia have been proposed. Amphetamines act to increase dopaminergic tone (enhanced sensitivity to dopamine) and are considered to produce some of the positive symptoms of schizophrenia (section 1.2) (Angrist *et al.,* 1974). These observations coupled with the knowledge that many of the neuroleptics that are effective in treating schizophrenia are antagonists of dopamine receptors (Scharfetter, 2001) have lead to the dopamine hypothesis of schizophrenia (Depatie and Lal, 2001 for review).

PCP and other noncompetative antagonists of NMDA receptors are considered to produce both positive and negative symptoms associated with schizophrenia and exasperate symptoms in schizophrenic patients (Javitt Zukin, 1991). These observations have led to the development of PCP and MK-801 (a highly specific noncompetitive antagonist of NMDA receptors) mouse models for schizophrenia (Carlsson and Svensson, 1990; Corbett *et al.,* 1993; Corbett *et al.,* 1995). These pharmacologically induced animal models of schizophrenia are of particular interest, not just because of their use in dissecting the neurochemical mechanisms of major mental illness, but also because they have allowed the development of behavioural paradigms for the assessment of schizophrenia models in non-humans (Crawley and Paylor, 1997 for review).

Reeler and *Scrambler* mouse mutants have been extensively investigated in terms of their behavioural and neuroanatomical phenotypes and the molecular basis of the mutations. The primary behavioural phenotypes of *Reeler* and *Scrambler* mice are those of impaired motor coordination, tremors and ataxia (D'Arcangelo *et al.,* 1995).

The neuroanatomical features of *Reeler* and *Scrambler* mutants share some parallels with malformation in the cerebral cortex of schizophrenics (section 1.2.3; Selemon *et al.,* 1998). Impagnatiello *et al.,* (1998) followed up this suggestive association with postmortem quantitation of the *RELN* gene (the gene mutated in *Reeler* mice) products in the brains of schizophrenics and matched controls. The finding of significantly reduced *RELN* mRNA and protein in the brains of schizophrenics compared to controls (approximately 50% reduction) has raised the profile of *RELN* as a candidate schizophrenia susceptibility locus (chromosome 7q22). These findings have yet to be independently supported or replicated. A further caveat is the disparity of phenotypes between those of *Reeler* mutants and the clinical presentation of schizophrenia, although the *Reeler* and *Scrambler* mice do show some of the behavioral traits of previously described mouse models for schizophrenia (D'Arcangelo *et al.,* 1995).

The *Dishevelled* homologue 1 (*DVL1*) gene when knocked out in mice resulted in fertile animals with no detected structural abnormalities (Lijam *et al.,* 1997). However, they displayed reduced social interaction (deficits in nest building, reduced huddling contact during sleep and subordinated responses in social dominance tests). Further investigation of these mice indicated that they possessed a sensorimotor gating deficit as measured by prepulse inhibition (see section 1.2.4) of acoustic and tactile startle responses. It was consequently suggested that *DVL1* null mice may provide a model for some aspects of the schizophrenia phenotype (Lijam *et al.,* 1997), particularly the negative symptoms. *DVL1* functions in the WNT signalling pathway that is involved in the development of several organs including the central nervous system.

A recent and very exciting mouse model of schizophrenia with reduced NMDA receptor expression has been reported by Mohn *et al.,* (1999). NMDA receptors are a subclass of iontropic glutamate receptors that are known to play an important role in long term potentiation and other aspects of synaptic plasticity. The NMDA receptors are hetromeric structures composed of an invariant NR1 subunit and one of four NR2 subunits (NR2A – NR2D). The different NR2 subunits determine channel properties

and their expression is strictly regulated in developmental time and space. There has been much speculation over the involvement of NMDA receptors in the molecular aetiology of schizophrenia (Javitt & Zukin, 1991; Coyle, 1996; Tamminga, 1998), the speculation arising largely from the pharmacological models of schizophrenia discussed previously.

Mice null for NMDA receptor components die perinatally (Forrest *et al.,* 1994; Kutsuwada *et al.,* 1996). Mohn *et al.,* (1999) generated a "leaky" knockout of the invariant NR1 gene, that expresses only 5% of wild type levels. These mice are able to survive to adulthood and display extensive behavioral similarities to the pharmacologically induced mouse models of schizophrenia (Corbett *et al.,* 1995 and references therein). Furthermore, the schizophrenia like symptoms can be alleviated with the use of neuroleptics effective in the treatment of schizophrenia. This mouse model has added considerable weight to the glutamate dysfunction hypothesis of schizophrenia, that was first implied by PCP induced psychosis. With the report of "smart mice" (increased learning ability) also being generated by transgenic manipulation of the NMDA receptor (Tang *et al.,* 1999; Bliss, 1999 for review) very rapid progress in this field is expected, gaining further insights into the roles of NMDA receptors in learning, memory and behaviour.

Although the diagnosis of schizophrenia in humans is problematic (section 1.3), the animal models discussed demonstrate that schizophrenia like traits can be modelled in animals. In addition to the battery of behavioural tests that have been developed to investigate animal models of schizophrenia, the models themselves will serve as useful references for future comparison with subsequently developed transgenic and knockout organisms.

1.5 The molecular basis for schizophrenia susceptibility

Due in part to the convincing evidence that schizophrenia has a substantial genetic contribution to susceptibility (section 1.3.4), many groups have attempted to identify the molecular basis of that genetic component. The strategies that have been used to determine schizophrenia susceptibility loci are whole genome scanning strategies,

candidate gene approaches or based on *a priori* implication of a locus (*i.e.* cosegregation of a previously mapped trait or cytogenetic aberration with schizophrenia). As yet, no mutation or polymorphism of a gene has been convincingly demonstrated to contribute to the genetic basis of schizophrenia. It is probable that attempts to define the genetic aetiology are confounded by the unclear mode of inheritance, partial penetrance, phenotypic heterogeneity and the likelihood that clinical schizophrenia is an umbrella for several aetiologically distinct diseases. Despite these hurdles, significant (section 1.5.3) results have been produced for several loci in independent data sets and in the hands of different researchers (table 1.2).

1.5.2 Positional cloning

Positional cloning is the use of naturally occurring DNA polymorphisms to identify regions of the genome that tend to be shared among affected relatives and tend to differ between affecteds and unaffecteds. The positional cloning strategy for the identification of trait or disease genes was first successfully employed in the cloning of the *CFTR* gene, the site of functional mutations leading to the disease cystic fibrosis (Riordan *et al.,* 1989). The general strategy used is to first establish linkage to a specific chromosome region and then linkage disequilibrium (LD) mapping to fine map the locus. Association studies using appropriately selected cases and controls may be required to further define the locus and identify the functional polymorphism.

1.5.3 Linkage studies

A linkage study identifies loci that segregate with the disease or trait of interest through a known pedigree. Consequently, linkage studies depend on one or more families with multiple affected individuals. The regions of a chromosome implicated by linkage can be further refined by LD mapping which defines recombination events between haplotypes. Recombination events provide a means of narrowing the region of interest by excluding regions of an original haplotype that fail to subsequently segregate with the trait. Technical issues have until recently (Douglas *et al.,* 2001) prevented the efficient direct detection of haplotypes, but in a defined pedigree the haplotypes can be confidently inferred from the phase of allele

transmission. Standard linkage analysis is useful for defining board regions of interest that are likely to contain a gene of major effect in a family or sub-set of families tested. Linkage analysis alone is insufficient to pinpoint a causative mutation or gene. Regions of the genome implicated by linkage analysis as containing schizophrenia susceptibility loci are summarised in table 1.2.

The results of linkage studies are evaluated in terms of a Z-score (non-parametric) or the logarithm of odds (LOD) score (parametric). Most commonly used is the LOD score which can take into account the mode of inheritance (recessive, dominant or additive), assumed recombination fraction (θ) and the disease spectrum (section 1.2). These parameters provide greater statistical power to the calculation but introduce problems of multiple testing. This is particularly the case for mental illness where the mode of inheritance is not known and the phenotypic spectrum of disease not well defined. The standard practice employed is to evaluate the linkage data under multiple models of disease and mode of inheritance and report the model with the most significant score (maximum LOD, MLOD), although the range of models investigated is typically reported (references in table 1.2). The most rigorous statistical correction for this multiple sampling is to multiply the observed *P*-values by the number of models tested (the Bonferroni correction). However, this correction is perceived as too conservative (Lander and Kruglyak, 1995) as the models are overlapping rather than independent. The issue of multiple testing is typically dealt with by simulation (Lindholm *et al.,* 2001) or importance sampling, a form of bootstrapping (Terwilliger and Ott, 1992; Ott 2001 for review).

A LOD score of 3 has been proposed (Risch, 1992) as a minimum threshold of significance in human genetic studies of monogenic conditions. A LOD of 3 means that the observed data is 10^3 fold more likely to arise under the specified hypothesis than under the null hypothesis of no linkage (odds ratio of 1000:1). Lander and Kruglyak, (1995) have argued that a LOD of 3 is insufficient to guard against false positive findings in whole genome studies of complex traits. Accordingly, guidelines of LOD >1.9 ($P=1.7\times10^{-3}$) and >3.3 ($P=4.9\times10^{-5}$) have been proposed as minimum thresholds for suggestive and significant linkage respectively (Lander and Kruglyak,

1995). Lander and Kruglyak, (1995) also proposed a criteria for "confirmed linkage" that requires a report of significant linkage and independent support of linkage at least to the nominal P value of 0.01. The observations of Lindholm *et al.,* (2001) based on computational simulation suggested that a LOD score of >2.2 should be considered suggestive and >3.7 should be considered significant.

There have been many linkage studies performed using markers spread across the whole genome. A comparable number of linkage and association studies restricted to areas of specific interest have been carried out, often as follow up studies (Baron 2001 for review). Positive reports of linkage have been made for almost every human chromosome. Coupled with non-replication of findings (Baron *et al.,* 1990; Owen *et al.,* 1990) and diminished significance on further scrutiny (McGuffin *et al.,* 1990) this field was viewed with great scepticism. However, the subsequent stricter adherence to the significance guidelines discussed above and larger sample sizes have resulted in statistically robust and replicated findings of "significant" and "confirmed" linkage (according to the criteria of Lander and Kruglyak, 1995). However, past experience appropriately dictates that these findings should still be treated with caution. Key findings of significant linkage and supporting findings of suggestive linkage for schizophrenia susceptibility loci are summarised in table 1.2. The findings of linkage in the chromosome 1q42 region (table 1.2) corresponds to the chromosome translocation breakpoint that is the focus of this work are discussed further in section 1.6.

OMIM[a]	Chromosome locus[b]	Score[c]	Study sample[d]	Reference[e]
604906	1q21-q22	6.50	Canada	Brzustowicz *et al.*, 2000
605210	1q43	4.34	Scotland	StClair *et al.*, 1990
	1q32-q44	3.82	Finland	Hovatta *et al.*, 1998
	1q33	3.60	Europe	Gurling *et al.*, 2001
	1q42	3.21	Finland	Ekelund *et al.*, 2001
	1q42.2	7.01	Scotland	Blackwood *et al.*, 2001
181510	5q11.2-q13.3	6.49	UK and Iceland	Sherrington *et al.*, 1988
603342	5q31	1.80	German and Israeli	Schwab *et al.*, 1997
	5q22-q31	3.35	Ireland	Straub *et al.*, 1997
	5q33	3.60	Europe	Gurling *et al.*, 2001
600511	6p23	3.90	Ireland	Wang *et al.*, 1995
	6p24-p22	3.51	Ireland	Straub *et al.*, 1995
	6p24-p22	2.20	Germany and Israel	Schwab *et al.*, 1995
	6p	2.19	Multiple	Moises *et al.*, 1995
603175	6q21-q22	3.06	US	Cao *et al.*, 1997
	6q21-q22	3.82	US and Australia	Martinez *et al.*, 1999
	6q25.2	7.70	Sweden	Lindholm *et al.*, 2001
603013	8p22-p21	3.64	US	Blouin *et al.*, 1998
	8p22-p21	3.48	Canada	Brzustowicz *et al.*, 1999
	8p21	P= 8e-6	Multiple	Pulver *et al.*, 2000
	8p21	3.60	Europe	Gurling *et al.*, 2001
603176	13q14-q32	4.18	US	Blouin *et al.*, 1998
	13q14-q32	4.42	Canada	Brzustowicz *et al.*, 1999
	13q14-q32	3.81	Canada	Brzustowicz *et al.*, 2000
NA	15q13-q14	5.30	US	Freedman *et al.*, 1997
	15q13-q14	3.57	US	Stober *et al.*, 2000

Table 1.2; Linkage studies in schizophrenia. Linkage studies that have found significant (as defined by Lander and Kruglyak, 1995; section 1.5.3) linkage or suggestive linkage that supports a report of significant linkage. Those findings that may reflect a common locus are clustered together. **(a)** The Online Mendelian Inheritance in Man (OMIM) reference number representing the described locus. Those loci that do not have an OMIM entry are indicated with NA. Some of the more recent references included in the table have not yet been incorporated into the corresponding OMIM entries. **(b)** The chromosome locus as defined in

the original report. **(c)** Maximum LOD score reported for the findings. 'P' indicates where a significant P-value was reported rather than a LOD score. **(d)** Geographical origin of the study sample. **(e)** Reference for the original findings.

1.5.4 Association studies

Association studies rely on similar principles as a linkage study. Whereas linkage studies test for the co-inheritance of loci between affected relatives, association studies test for the association of an allele with disease. The most frequently used method is the case-control association study which is based on the comparison of allele frequencies in a group of cases versus a group of controls. It is important that the control set is well matched for geographical, ethnic and social background or any positive findings may be due to population stratification rather than disease association (Lander and Kruglyak, 1995). An over-representation of an allele in the disease set relative to the control set suggests that that allele may be in linkage disequilibrium with the causative mutation. A variation on the case-control association study is the transmission disequilibrium test (TdT). TdT uses trios of an affected sibling, an affected parent and an unaffected parent and tests for a bias of allele transmission with disease. This study design incorporates the unaffected parent as a well matched control. Unlike linkage studies, association studies have considerable power to detect genes of small effect, but require large data sets to produce statistically significant results.

To date, association studies have been used to follow up studies of linkage findings and the testing of functional candidates for association with disease. A lack of evenly spaced polymorphic markers across the genome has previously prevented the use of whole genome association studies to identify susceptibility loci for major mental illnesses. Developments in genotyping large numbers of markers in parallel (Taylor, 2001a for review) along with the impending release of approximately 0.3 million ordered single nucleotide polymorphisms across the human genome (http://snp.cshl.org/) suggest that such studies can soon commence.

The rationale behind selection of functional candidate genes has been wide ranging, including genes of neurodevelopmental, immunological and neural signalling

importance. Much attention has been focused on the genes whose products are known to interact with the neuroleptics used to treat mental illness (O'Donovan and Owen, 1999 for review). Association studies of genes such as the 5HT2a seratonin receptor gene (Williams *et al.,* 1997) and the dopamine D3 receptor gene (Williams *et al.,* 1998) have resulted in suggestive association, but with insufficient significance to reject the hypothesis of no association. In contrast, Wei and Hemmings (2000) reported a highly significant (P=0.000036) association between the *NOTCH4* locus and schizophrenia. The importance of this finding was further emphasised by the robustly supported linkage evidence for a schizophrenia susceptibility locus over the region of chromosome 6p that contains *NOTCH4* (table 1.2). However, several attempts to replicate the finding of significant association have failed (Ujike *et al.,* 2001; McGinnis *et al.,* 2001; Sklar *et al.,* 2001; Imai etal., 2001). As yet, there has not been a replicated significant association of schizophrenia with any locus.

1.6 The t(1;11) family

The MRC Cytogenetics Registry (Edinburgh, UK) contained karyotypic and clinical data on 282 human pedigrees segregating autosomal anomalies. In a survey of this resource, StClair *et al.,* (1990) identified a single large pedigree that co-segregated a reciprocal translocation between the long arms of chromosome 1 and chromosome 11 with a spectrum of major mental illnesses including schizophrenia, schizoaffective disorder, recurrent major depression and adolescent conduct disorders (maximum LOD = 4.34). Cytogenetic analysis suggested that the translocation was between the q43 region of chromosome 1 and q21 on chromosome 11 (StClair *et al.,* 1990) although this has since been refined to 1q42.2 and 11q14.3 (Millar *et al.,* 2000). Within the MRC Cytogenetics Registry and in early papers (Jacobs *et al.,* 1970; StClair *et al.,* 1990; Fletcher *et al.,* 1993) the family was referred to as family "K26", in subsequent publications they have been referred to as the "t(1;11) family", this later terminology will be adopted for the remainder of the thesis.

The t(1;11) family was first identified and recorded by the MRC Cytogenetics Registry during a cytogenetic survey of mental hospitals and borstals (Jacobs *et al.*, 1970). A physically normal 18 year old male was found to have an apparently balanced translocation between chromosome 1 and a C-group chromosome (subsequently identified as chromosome 11; figure 1.1). Cytogenetic investigations of the relatives of this individual revealed three cytogenetic abnormalities segregating in the family, a Robertsonian translocation between two D-group chromosomes, the t(1;11) translocation and an "unusually large secondary restriction" on chromosome 1, proximal to the t(1;11) breakpoint (Jacobs *et al.*, 1970). No physical abnormalities were reported to be coincident with the cytogenetic abnormalities in the family. Neither the Robertsonian translocation nor the secondary restriction were observed to segregate with mental illness in the family (StClair *et al.*, 1990).

Balanced translocations in an individual often result in disrupted chromosome disjunction during meiosis and the subsequent production of genetically unbalanced embryos, a scenario typified by familial Downs syndrome (OMIM: 190685). The absence of such unbalanced embryos in the t(1;11) family (Jacobs *et al.*, 1970) suggests that there is either a meiotic mechanism preventing the production of genetically deficient or overdosed gametes, or more likely the genetic imbalance results in early embryonic lethality. The co-segregation of the t(1;11) translocation and mental illness was not observed in these early studies on the family.

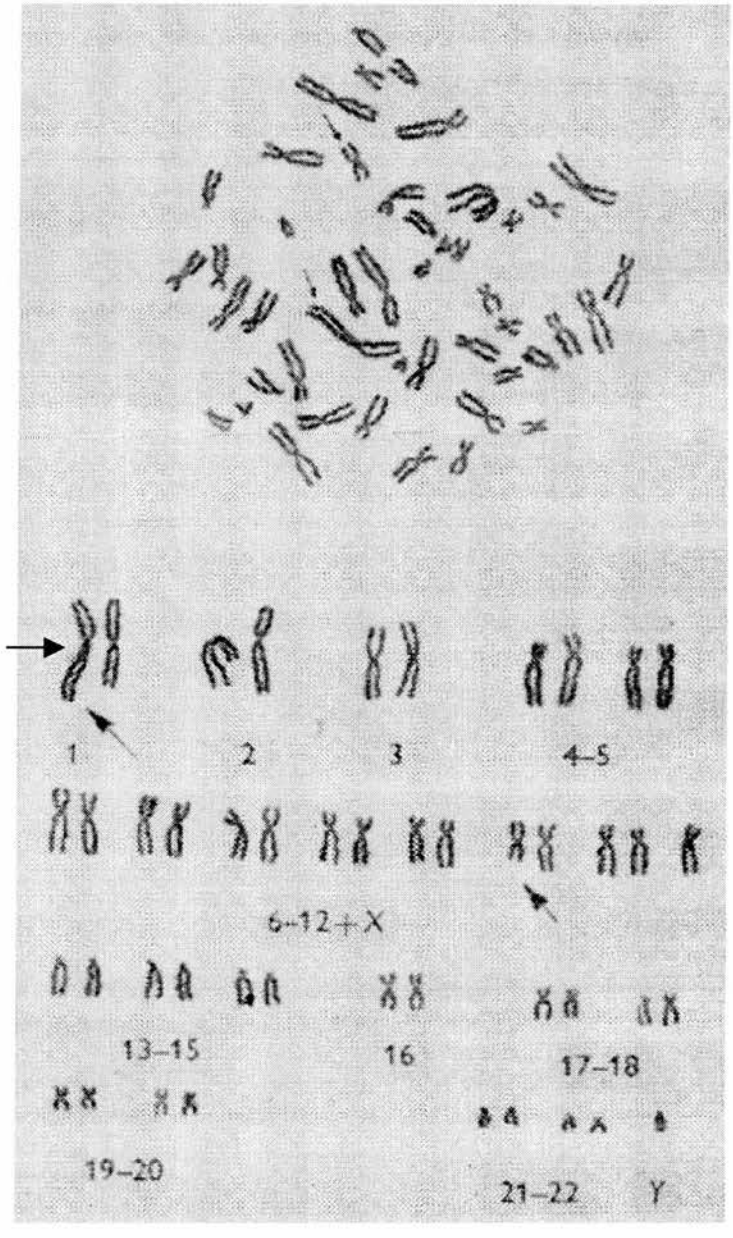


Figure 1.1; Karyotype of the proband individual from the t(1;11) family. Reproduced with permission P. Jacobs and Cambridge University Press; from Jacobs *et al.,* 1970. Arrows indicate the reciprocal translocation between chromosome 1 and the C – group chromosome subsequently identified as chromosome 11. In this individual the unusually large secondary restriction (horizontal arrow) is in *cis* with the derived chromosome 1. The secondary restriction was not observed to segregate with mental illness in the family, or with any other phenotype.

In the original search for an association between the t(1;11) translocation and mental illness, 34 of the 77 family members available for cytogenetic analysis were found to carry the t(1;11) translocation (StClair *et al.,* 1990). Of the 34 translocation carriers, 16 had a psychiatric diagnosis whereas only five family members without the translocation had a psychiatric diagnosis, and none of these were major mental illnesses (three alcoholism, one minor depressive disorder and one generalised anxiety). As previously discussed, the spectrum of aetiologically related psychiatric phenotypes is not well defined (section 1.2). For this reason the significance of the association was tested under several diagnostic definitions, the caveats of this approach relating to multiple testing are discussed in section 1.5.3. Under a diagnostic definition restricted to schizophrenia and schizoaffective disorder, a maximum LOD of 2.19 was observed. Including recurrent major depression in the phenotype increased the maximum LOD to 3.33. Adding adolescent conduct and emotional disorder increased the maximum LOD to 4.34. Further inclusion of minor psychiatric phenotypes such as generalised anxiety, minor depressive episode and chronic alcoholism substantially reduced the maximum LOD to 1.49 (StClair *et al.,* 1990).

Subsequent clinical and cytogenetic follow up of the t(1;11) family has further supported the association of major mental illness and the t(1;11) translocation. In a systematic follow up over a ten year period, original and new members of the family were interviewed by psychiatrists and case notes reviewed using the DSM-IV (American Psychiatric Association, 1994) standardised diagnostic criteria (Blackwood *et al.,* 2001). It is important to note that both the primary psychiatric diagnoses (St Clair *et al.,* 1990) and the follow up diagnoses (Blackwood *et al.,* 2001) were carried out by psychiatrists blind to karyotype status. As a result of the follow up study, nine new cases of major mental illness were diagnosed and ten new karyotypes were obtained, predominantly from a fifth generation of the family who were too young to be included in the original study (St Clair *et al.,* 1990).

A total of 87 members of the family have been karyotyped. Of these 37 were found to carry the translocation. The karyotype status and psychiatric diagnosis of the family is summarised in figure 1.2. Of the 37 translocation carriers, 18 (49%) had a major psychiatric diagnosis (schizophrenia, bipolar disorder or recurrent major depression). No major psychiatric disorder was diagnosed in family members without the t(1;11) translocation. Using the data set of the follow up study, a maximum LOD score of 7.1 was obtained based on a clinical phenotype of schizophrenia, bipolar disorder and recurrent major depression (Blackwood *et al.,* 2001). Significant linkage was also demonstrated between the t(1;11) translocation and a clinical phenotype restricted to just schizophrenia (LOD score 4.5) or just bipolar affective disorder and recurrent major depression (LOD score 3.6).

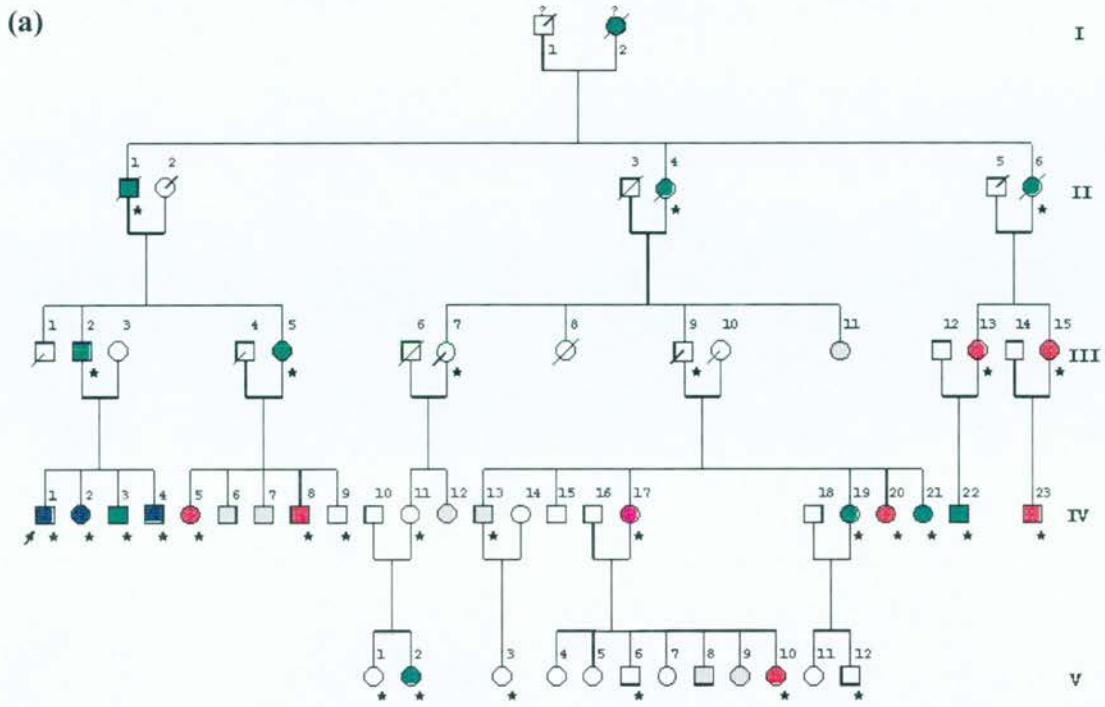


Figure 1.2; Co-segregation of a balanced translocation with major mental illness. **(a)** Branches of the t(1;11) family that segregate the t(1;11) translocation. Asterisks indicate the t(1;11) carriers. An arrow indicates the proband individual whose karyotype is shown in figure 1.1. Major psychiatric diagnosis: schizophrenia or schizoaffective disorder (red); bipolar disorder (pink); recurrent major depression (green). Other psychiatric diagnosis: adolescent conduct disorder (blue); generalised anxiety, alcoholism and minor depression (grey). The karyotypes for generation I are unknown. **(b)** Ideograms of normal chromosomes 1 (black) and 11 (red) and the derived (Der) chromosomes 1 and 11. The cytogenetic breakpoint position is 1q42.1 and 11q14.3.

In addition to psychiatric interviews during the follow up study of Blackwood *et al.,* (2001), family members were evaluated for physical dysmorphisms, intelligence quotient (IQ) and auditory P300 event related potential (ERP) (section 1.2.4). Translocation carriers, non-translocation family members and controls showed IQ's in the normal range, there was no association of karyotype status and physical dysmorphism (Blackwood *et al.,* 2001). The reduced amplitude and prolonged latency of P300 ERP has been proposed as a biological marker of susceptibility to major mental illness (Schreiber *et al.,* 1992; Blackwood *et al.,* 1999; see section 1.2.4). To evaluate the P300 ERP in the t(1;11) family; 12 t(1;11) carriers, 10 family members without the translocation, 20 unrelated schizophrenics and 26 unrelated controls were compared. There was no significant difference in P300 ERP measures between t(1;11) carriers with and without psychiatric diagnoses. Further, the P300 ERP results were similar between translocation carriers and the unrelated schizophrenics but significantly different (>2 standard deviations from the mean) from the family members without the translocation and the unrelated control group (Blackwood *et al.,* 2001). These findings are consistent with there being a highly penetrant underlying neuropathology in t(1;11) carriers that manifests as major mental illness only 50% of the time.

It has been proposed that psychiatric disorders represent complex aetiological interactions between multiple genes and the environment (section 1.3.5). However, within the t(1;11) family, major mental illness appears to be segregating as an autosomal dominant trait exhibiting partial (approximately 50%) penatrance (figure 1.2). The highly significant correlation (LOD score 7.1) of the t(1;11) translocation with the major mental illness in this family has lead to the following hypothesis. A gene or genes located at or near the breakpoint on chromosome 1 or 11 is a gene of major effect in conferring susceptibility to mental illness in this family. Conceivably the gene could be directly disrupted by the breakpoint, its regulation influenced by the breakpoint or a mutation unrelated to the breakpoint is recombinatorially fixed by the breakpoint.

It is a particular advantage of using cytogenetic markers in disease association that the cytogenetic aberration itself is a good candidate for the causative mutation. Hypothetically, a gene directly disrupted by a translocation breakpoint would be positionally the strongest candidate, followed by neighbouring genes within the range of position effects (section 1.6.2). A causative mutation that has been recombinatorially fixed by the suppression of recombination around a translocation breakpoint (section 1.6.3) would represent a worst case scenario for defining the molecular nature of increased susceptibility to major mental illness. Each of these possibilities is considered in turn.

1.6.1 A gene directly disrupted by the breakpoint

A gene is considered directly disrupted if its transcription normally proceeded across the position of the translocation breakpoint with exons or untranslated transcribed sequence located on either side of the breakpoint. If a gene is directly disrupted, it may no longer make a product, a fusion product could be produced or a truncated product could be produced. The apparent dominant mode of inheritance in the t(1;11) family, could be explained by haplo-insufficiency, if a functional gene product was not produced from the translocation chromosome, or gain of function (or loss of function suppression) by truncation or gene product fusion. Gain of function through chromosome translocation mediated gene fusion events is a well characterised phenomenon in neoplastic progression (Lengauer *et al.,* 1998 for review).

If a gene or genes were directly disrupted by the translocation breakpoints, functional and structural assessment of the genes could give insight into the influence of the translocation on their function. Transcript and protein levels could also be directly tested for. Truncation or fusion gene products could be identified at the RNA or protein level, although obtaining appropriate tissues from translocation carriers would present a problem. Lympoblastoid cell lines from translocation carriers have been made (Fletcher *et al.,* 1993) and could be used in some circumstances for this type of investigation.

1.6.2 Gene regulation disrupted by the breakpoint

There have been several examples were chromosomal translocations or inversions in humans lead to disease, where the transcription unit of the causative gene is not directly disrupted by the breakpoint. In all cases the transcriptional regulation of these genes was profoundly affected (Kleinjan and van Heyningen, 1998 for review). Such position effects are caused by the removal and/or addition of transcriptional enhancers and suppressors in relation to the transcriptional start site of the gene, or through the influence of altered local chromatin environment. The range of such influences can vary considerably between loci. Chromosomal breakpoints associated with aniridia (OMIM:106200) have been reported 125 kb from the *PAX6* gene responsible for the phenotype, and in X-linked deafness (OMIM:300039) micro-deletions have been found to be 900 kb from the *POU3F4* gene responsible for the phenotype (deKok *et al.,* 1996). A transgene insertion in mouse suggests that a *cis* regulator of Sonic hedgehog is located at least 1 Mb from the transcription unit of the gene (S Heaney, personal communication). The influence of alterations to local chromatin environment could potentially be active over even greater distances (Walters *et al.,* 2001).

1.6.3 Linkage with the causative mutation

It has been well established in a variety of organisms that chromosomal rearrangements, including reciprocal translocations, can cause a suppression of recombination in the proximity of the rearrangement (Dobzhansky and Sturtevant, 1931; Davisson and Akeson, 1993). Davisson and Akeson, (1993) have reported that recombination was suppressed in regions of up to 40 cM around Robertsonian translocations in mice. However in the same study, another Robertsonian translocation did not show any significant suppression of recombination.

Two scenarios could give rise to a causative mutation in *cis* with the t(1;11) translocation: The translocation involves a chromosome that already contains a mutation that increases susceptibility to major mental illness. Or, the translocation occurs and a mutation subsequently arises on one of the translocation derived chromosomes. In either case, if the true causative mutation was physically close to

the translocation breakpoint it would be in linkage disequilibrium with the breakpoint. However, if the tranlocation suppressed recombination around the breakpoint, the extent of linkage disequilibirum could be in the range of 10's of megabases in physical distance, encompassing hundreds of genes.

This issue has previously been addressed in the t(1;11) family. He *et al.,* (1996) typed polymorphic markers within 10 cM of the translocation breakpoint on both chromosomes 1 and 11. Although the statistical power of the study was limited, no significant reduction ($p = 0.1$) in recombination frequency was observed around the breakpoint on chromosomes 1 or 11. Observed recombination events also defined a 20 Mb (based on marker positions within the August 12th 2001 Golden Path human genome assembly, section 2.11.1) maximum region of chromosome 1 by haplotype analysis.

It is doubtful that a causative mutation linked to the t(1;11) translocation could be identified using the t(1;11) family as the only resource. The elucidation of such a causative mutation would require the use of multiple pedigrees showing linkage to the same region.

1.6.4 Validation of findings

The reciprocal chromosome 1 - 11 translocation that occurred in an ancestor of the t(1;11) family is likely to be a unique event whose consequences may not be readily recapitulated by more common point mutation and small insertion/deletion mechanisms. Functional investigations of genes at and near the breakpoints, and animal models will be important to assess the candidacy of genes for a contribution to the susceptibility of major mental illness. However, validation of a susceptibility gene for mental illness would require the identification of a causal mutation within the t(1;11) family and in one or more families that do not segregate the t(1;11) translocation. For this reason the recent findings of significant linkage to schizophrenia in chromosomal regions encompassing the chromosome 1 breakpoint locus are extremely important (Hovatta *et al.,* 1998; Ekelund *et al.,* 2001; Gurling *et al.,* 2001).

Hovatta *et al.,* (1998) reported significant linkage (LOD score of 3.82) in the large interval from 1q32 to q44 in a population based Finish family study. In a subsequent follow up study, the same group genotyped 147 microsatellite markers over a 45 cM region of the q arm of chromosome 1. A total of 557 affected (schizophrenia under DSM-IV criteria or schizophrenia spectrum mental illness) individuals from 221 families were genotyped. The strongest evidence for linkage (LOD = 3.21) was to the marker D1S2709 (Ekelund *et al.,* 2001) which is located only 80 kb teleomeric to the chromosome 1 breakpoint and is intragenic to the protein coding gene directly disrupted by the breakpoint (section 1.6).

In an independent linkage study using families from across Europe, Gurling *et al.,* (2001) identified LOD scores of >3.0 at five loci in the genome including 1q33 at a maximum LOD score of 3.2. Although the peak of linkage was centromeric to the breakpoint, the broad region of linkage detected around the 1q33 locus was considered supportive of a schizophrenia susceptibility locus in the distal region of chromosome 1q (Gurling *et al.,* 2001), including the breakpoint locus.

The finding of other families that show linkage for a schizophrenia susceptibility locus to the same region as the chromosome 1 breakpoint, provide the opportunity to pinpoint genes and causal mutations. These observations suggest that the molecular basis of disease in the t(1;11) family is not unique and may be of substantial epidemiological significance. With the independent validation of the chromosome 1 locus, 1q42 now represents one of the best, if not the best candidate locus for a schizophrenia susceptibility gene.

1.6.5 The chromosome 11 breakpoint

In addition to the t(1;11) family, two other balanced chromosomal translocations involving the long arm of chromosome 11 had been identified (Smith *et al.,* 1989; Holland and Gosden, 1990) that appeared to co-segregate translocation chromosomes with major mental illnesses. Although the co-segregation of major mental illness was not as clear cut as in the t(1;11) family, it did highlight the chromosome 11 region as an area of particular interest. Also, the tyrosinase gene and the dopamine receptor D2 both mapped to the general region of 11q represented by the t(1;11) breakpoint

(Barton *et al.,* 1988; Grandy *et al.,* 1989). Tyrosinase was of particular interest as a large family had been reported to co-segregate tyrosinase deficient albinism and schizophrenia spectrum mental illness (Barton *et al.,* 1988). Dopamine receptor D2 was of interest as it is a neural receptor with a high binding affinity for a range of antipsychotic drugs. For these reasons, the chromosome 11 breakpoint region was argued to be a better candidate region than the chromosome 1 breakpoint (StClair *et al.,* 1990; Fletcher *et al.,* 1993; Evans *et al.,* 1995) and was consequently investigated before the 1q42 region.

A contiguous yeast artificial chromosome (YAC) map was constructed over 3 Mb of the chromosome 11 breakpoint region and was shown to traverse the site of the breakpoint (Evans *et al.,* 1995). Using the clone map as a primary resource, expressed sequences were identified from the region around the chromosome 11 breakpoint. An alpha-tubulin pseudogene approximately 250 kb from the breakpoint and a cluster of novel cDNA fragments mapping approximately 700 kb from the breakpoint were identified (Devon *et al.,* 1997). No genes were found to be directly disrupted by the breakpoint on chromosome 11 (Devon *et al.,* 1997; Semple *et al.,* 2001). However, three functionally notable genes were identified within a 2.51 Mb fragmented sequence contig around the breakpoint reigon, Naaladase II, its close homologue *FOLH2* and the glutamate receptor *GRM5* (Semple *et al.,* 2001). Naaladase II and by inference through homology *FOLH2*, both hydrolyse N-acetyl-L-aspartyl-L-glutamate (NAAG), a storage form of the neurotransmitter glutamate (Pangalos *et al.,* 1999). *GRM5* is the gene encoding the metabotrophic glutamate receptor 5. All three of these genes are rational functional candidates for involvement in the susceptibility of major mental illness (Semple *et al.,* 2001) and their genetic investigation as functional and positional candidates in both the t(1;11) family and case – control association studies is underway (Devon *et al.,* 2001; P. Thompson, personal communication).

1.6.6 The chromosome 1 breakpoint

A restriction fragment containing the chromosome 11 breakpoint was used to identify a translocation breakpoint containing clone from an *EcoRI* genomic library constructed, from a translocation carrier derived cell line (Millar *et al.,* 2000).

Sequence from the breakpoint containing clone was then used to construct a cosmid and PAC clone contig across the chromosome 1 region (Millar *et al.*, 2000a). Sequencing of a 7.3 kb EcoRI fragment containing the chromosome 1 breakpoint lead to the identification of a novel, potentially protein coding gene subsequently referred to as Disrupted In SChizophrenia 1 (*DISC1*). Evidence was also found for a second transcript that was potentially transcribed from the opposite strand to *DISC1*. This transcript is subsequently referred to as Disrupted In SChizophrenia 2 (*DISC2*).

The wider genomic context and structure of the *DISC1* and *DISC2* genes was unknown. The sequence obtained for *DISC1* suggested that it had the potential to encode a protein, but a lack of sequence homology or recognised functional motifs excluded functional predictions. There was no obvious protein coding open reading frame in the *DISC2* transcript. It was thought likely that the *DISC2* transcript represented the 3' UTR of an as yet unidentified protein coding gene. Multiple expressed sequence tags (ESTs) had also been mapped onto the clone contig through sequence similarity or hybridisation (summarised in section 6.2.1). The significance of these EST sequences was unknown. It is the molecular and comparative investigation of this locus and the flanking genes that is the principal focus of the work presented in subsequent chapters.

1.7 Experimental approaches

At the time of starting this work there was no genomic sequence in the public sequence databases (section 2.11.1) that could be identified as being from the chromosome 1 breakpoint region. However, this work was carried out concurrently with that of the International Human Genome Sequencing Consortium (IHGSC). It was therefore expected that substantial amounts of sequence representing all regions of the genome, including the chromosome 1 breakpoint would be generated. This sequence could allow gene structures to be accurately resolved by alignment of cDNA sequences to genomic sequence. Flanking genes might be identified by gene prediction and homology to known genes, rather than the speculative hybridisation of ESTs onto clone contigs, low coverage sample sequencing and cDNA selection strategies that would have previously been used. As well as allowing the rapid

preliminary annotation of the genomic region around the chromosome 1 breakpoint, genomic sequence could serve as a platform for the detailed identification of other functionally significant features that may not be protein coding, such as transcriptional regulatory elements and non-protein coding transcripts.

Currently available *ab initio* gene prediction software accurately predicts approximately 64% of protein coding exons with a specificity of 44% when provided with genomic sequence containing more than one gene (Gugio *et al.,* 2001). A consensus approach integrating the results of multiple prediction methods can be used to increase specificity at the expense of sensitivity (Venter *et al.,* 2001). Considering the complexity of the problem these prediction accuracies are impressive from a computational perspective, but are inadequate for biological investigation of function and mutation screening.

A typical approach that is also used later in this work, is to take the best consensus exon predictions integrated with EST sequence similarity and protein homology, and perform RT-PCR between them. While this approach can be extremely effective (it was the approach used to identify the *DISC1* transcript, Millar *et al.,* 2000a), it is reliant on the *ab initio* prediction of an exon. If an exon was not consistently predicted and was not represented by EST sequence or protein homology it would not be detected. A medically relevant example of this scenario is that of the *RPGR* gene. Mutations in the *RPGR* locus were shown to be responsible for a form of retinitis peigmentosa (OMIM: 312610), but no mutations could be found in up to 60% of the patients predicted from linkage studies. Analysis of the genomic sequence and comparison to *Fugu* genomic sequence lead to the discovery of a previously unknown, alternatively spliced terminal exon (Vervoort *et al.,* 2000). This exon was subsequently shown to contain the "missing" mutations (Vervoort *et al.,* 2000).

1.7.1 Model genomes

Ideally for comparative genomic analysis, contiguous sequence over the region of interest would be available from a range of species including relatively close and distantly related species. Closely related species (for example, mammal to mammal)

are likely to have a high degree of sequence identity within coding and other functionally conserved regions. However, there is also a high degree of conservation that does not appear to reflect functional conservation of sequence. It can therefore be problematic to distinguish functionally relevant conservation from sequences that have simply not had sufficient evolutionary time to accumulate random changes (Hardison *et al.,* 1997 for review). Recently developed techniques such as Waba (Kent, in press), Twinscan (Korf *et al.,* 2001) and Rosetta (Batzoglou *et al.,* 2000) have started to address this issue from the perspective of coding sequence prediction.

More distantly related species comparisons, such as mammal to bird, amphibian or fish, show poorer conservation of coding and regulatory sequences. However, the signal to noise ratio is substantially improved. During the approximately 350 million years since the last common ancestor of mammals and fish, nucleotide substitutions, insertions and deletions have eliminated any sequence similarity that has not been constrained by evolution (with the exception of low complexity repeats). Consequently, it is hypothesised that any conservation of sequence in orthologous regions is due to functional constraint.

Fugu rubripes has been proposed as a model vertebrate genome for use in comparative genomic studies, particularly for the annotation of the human genome (Brenner *et al.,* 1993). *Fugu* is a member of the Tetraodontidae sub-family of fish (pufferfish). Tetraodontiforms have a haploid genome of approximately 400 Mb, compared to the 3,000 Mb of the human haploid genome (Elgar, 1996). *Fugu* has also been shown to have a similar gene repertoire to other vertebrates (Elgar *et al.,* 1999; Crollius *et al.,* 2000; McLysaght *et al.,* 2000). This 7.5 fold genomic compaction in *Fugu* relative to humans reflects a reduced amount of repetitive sequence and a consequently higher gene density. The coding sequence of *Fugu* genes is typically of the same length as other vertebrates but intergenic and intragenic spaces are typically smaller (Brenner *et al.,* 1993; Elgar *et al.,* 1996; McLysaght *et al.,* 2000).

As proposed by Brenner *et al.,* (1993), the compact, largely repeat free genomic sequence of *Fugu* has been successfully used to find new genes, predict gene structures, identify previously unknown exons and define conserved non-coding sequences. Comparative genomic studies using *Fugu* include the Huntington disease gene (Baxendale *et al.,* 1995), the WAGR region (Miles *et al.,* 1998), the HOX clusters (Aparicio *et al.,* 1997) and the Surfit gene cluster (Armes *et al.,* 1997) (for comprehensive review see McLysaght *et al.,* 2000). Within these regions of the *Fugu* genome there was generally conserved synteny with a good correlation of gene order and orientation with the orthologous region in humans. An exception was the Surfit gene cluster, which had undergone substantial, local gene order rearrangement. In a study by McLysaght *et al.,* (2000), it was estimated that 40 to 50% of gene pairs have conserved synteny with a human chromosome, although this is likely to be a substantial underestimate considering the incomplete nature of human genomic sequence used.

The utility of *Fugu* as a model genome for comparative analysis with mammalian sequences is complemented by the availability of *Fugu* genomic and cDNA libraries. *Fugu* was therefore an obvious choice for the comparative sequence analysis of the human chromosome 1 breakpoint region.

1.8 Aims

The aims of this research were to identify and characterise the genes around the chromosome 1 translocation breakpoint. A combination of molecular biological and comparative genomic approaches were used to investigate this locus. The identified genes and other sequence features were evaluated for their potential disruption by the t(1;11) translocation and functional candidacy for a role in the pathology of t(1;11) carriers. Specifically the aims were:

- Identify, isolate and sequence the chromosome 1 breakpoint orthologous region from a model vertebrate genome, *Fugu rubripes.*
- Refine the transcription map of the human breakpoint locus.
- Construct a transcription map for *Fugu.*
- Compare the genomic sequence and transcription maps of the human and model genomes.
- Investigate the function and conservation of proteins encoded near the chromosome 1 breakpoint.
- Investigate the nature of the *DISC2* transcript.
- Evaluate the candidacy of genes at the chromosome 1 translocation breakpoint for involvement in the susceptibility to major mental illness.

Chapter 2

Methods

2.1 Preface

This chapter in conjunction with appendices I and II, contains the technical information necessary to replicate the findings of all experiments discussed in subsequent chapters. Standardised protocols used with minimum modification are discussed in brief, citing the original protocol. Protocols that have been devised specifically for this work, or modified significantly from the original protocol are discussed and a complete protocol provided. Unless otherwise stated, protocols were developed form Sambrook *et al.,* (1989).

Software and datasets used in the computational aspects of the work are also documented in this chapter. Software version number, default settings and any modifications to the software are described. Custom software developed during this work is also documented and complete source code included in appendix II.

2.2 Bacterial cell culture and DNA

2.2.1 Solutions

Luria – Bertani broth (LB)

1% (w/v) Bacto-Tryptone (Difco); 0.5% (w/v) Bacto-Yeast extract (Difco); 0.1% (w/v) NaCl; pH adjusted to 7.0 with NaOH. Sterilized by autoclaving.

Terrificbroth (TB)

1.2% (w/v) Bacto-Tryptone; 2.4% (w/v) Bacto-Yeast extract; 0.4% (v/v) glycerol; After autoclaving add (17 mM KH_2PO_4; 72 mM K_2HPO_4).

P1

50 mM Tris.Cl, pH 8.0; 10 mM EDTA; 100 µg/ml RNaseA (Roche).

P2

200 mM NaOH; 1% (w/v) SDS.

P3

3 M potassium acetate, pH 5.5.

Neutralizer

1 M Tris.HCl; 2 M NaCl, pH 5.5.

Denaturer

0.5 M NaOH; 1.5 M NaCl.

2.2.2 Bacterial cell culture

The major bacterial (*E.coli*) strains used through this work were XL1-Blue (Bullock *et al.,* 1987) and TOP10 (for genotypic description see Invitrogen TOPO TA Cloning Kit, version E2). Both strains were found to grow well in LB broth and TB. Bacteria were grown at 37°C unless otherwise stated. Typically, liquid phase bacterial growth for volumes in excess of 10 ml was in conical flasks of at least 5× the volume of liquid media. Culture volumes of 2 to 10 ml were grown in 50 ml plastic tubes. Liquid media cultures were rotationally agitated at 200 rpm for the duration of their growth, optimally 12 hours. For cultures in excess of 100 ml, an overnight culture of 1% volume was used as a starter culture.

All culture of bacteria on solid media was carried out on LB plates (LB broth with 15 g of Bacto-Agar per litre of liquid media) with appropriate selective antibiotics (section 2.2.3).

For long term storage of bacteria, 0.8 ml of overnight culture was mixed with 0.2 ml of 50% glycerol and stored at -70°C. In 96 well format, 80 µl of culture was mixed

with 20 µl of 50% glycerol. In a final concentration of 20% glycerol at -70°C *E. coli* was found to remain viable for several years.

2.2.3 Bacterial selection and screening

Ampicillin (Amp) was used at a working concentration of 50 µg/ml. Kanamycin (Kan) was used at a working concentration of 30 µg/ml. Chloramphenicol (Cm) was used at a working concentration of 100 µg/ml or 30 µg/ml when RPCI-23 BAC clones (section 2.2.4) were being grown. Where blue/white colour selection was used, LB-agar plates contain final concentrations of 0.2 mM IPTG and 40 µg/ml X-gal. All of the antibiotics, IPTG and X-gal used were supplied by Sigma.

2.2.4 Vectors

Table 2.1 summarises vectors used during this work, their principal properties and context of use.

Vector[a]	Type[b]	Selective marker[c]	Colour selection[d]	Context of use[e]
pBluescript SK-	Plasmid	Amp	Yes	*Fugu* cDNA libraries. Sticky ended cloning.
pCR2.1-TOPO	Plasmid	Amp + Kan	Yes	PCR product cloning.
pCR4.0-TOPO	Plasmid	Amp + Kan	Yes	PCR product cloning.
Lawrist 4	Cosmid	Kan	Yes*	*Fugu* Genomic DNA library.
pBeloBACII	BAC	Cm	Yes*	*Fugu* Genomic DNA library. Mouse RPCI-23 genomic BAC library.

Table 2.1; Bacterial cloning vectors. **(a)** The name of the vector. **(b)** The type / family of vectors. BAC stands for bacterial artificial chromosome. **(c)** Selective marker indicates which antibiotics the vector confers resistance to. **(d)** Insertion into the cloning site disrupts expression of the *LacZ* gene causing bacterial colonies to be white instead of blue when grown on media containing X-gal (section 2.2.3). Asterisks indicate that although colour selection was available in these vectors it was not used in the work described. **(e)** Brief description of the application of each vector.

2.2.5 Alkaline lysis

The alkaline lysis method of nucleic acid preparation was common to all of the subsequent protocols in section 2.2. The principal aim was to isolate vector DNA from genomic DNA and all other bacterial cell debris. For the compositional definition of buffers, see section 2.2.2.

A pellet of bacterial cells was resuspended in buffer P1 so that no clumps of cells remained. Buffer P2 was added and mixed gently by inversion. The preparation was incubated at room temperature for 5 minutes and subsequently on ice for 10 to 30 minutes until a consistent viscous solution was formed. During this incubation, bacterial cells were lysed and RNA was degraded by the RNase A present in P1. NaOH present in P2 causes substantial denaturation of nucleic acids and proteins released from lysed cells. Following incubation, buffer P3 was added and gently mixed, causing the rapid precipitation of cell debris, SDS and genomic DNA as large white/grey aggregates. Episomal DNA (such as plasmids, cosmids and BACs) remain largely in solution, protected from denaturation by its relatively small size and compact supercoiled nature. Vector DNA was subsequently separated from genomic DNA and other cell debris by centrifugation (10,000 g for 30 minutes).

Typically, 10 µl of each P1, P2 and P3 was used per 1000 µl of bacterial culture. Plasmid preparations were propan-2-ol precipitated (section 2.4.2) and washed in 75% ethanol (section 2.4.2), after which they were of sufficient purity to analyse by restriction digestion or DNA sequencing. Larger vectors such as cosmids and BACs required further purification (sections 2.2.6, 2.2.7 or 2.4.3) before they could be reliably used for digestion or sequencing. Further purification steps also substantially increased the shelf life of all preparations.

2.2.6 Caesium chloride gradient centrifugation

This protocol when used for the purification of plasmid or cosmid DNA, was performed exactly as described in Sambrook et al (1989). The $CsCl_2$ gradient purification protocol was found to yield DNA of a similar purity to the ion exchange

protocol (section 2.2.7). Due to the toxic and labour intensive nature of this procedure it was only rarely used.

2.2.7 Ion exchange purification

Supernatant from alkaline lysis was cleaned of contaminants using columns of ion exchange resin. Supernatant was added to a column under salt conditions optimal for binding of DNA, the column was washed removing contaminants and DNA subsequently eluted by adjusting the salt conditions ("Qiagen plasmid kit handbook" (QPKH) for complete protocol and formulation of supplied buffers).

The low copy number and large size of cosmid and BAC vectors necessitated modifications to the standard Qiagen protocol. The initial volume of alkaline lysis supernatant was too great and DNA concentration sub-optimal for a Qiagen-tip™ ion exchange column. DNA was concentrated by propan-2-ol precipitation and ethanol washing (sections 2.4.2 and 2.4.1) of the alkaline lysis supernatant. Precipitated DNA was resuspended with TE (pH 7.0) to 0.1× the original volume. To provide optimal conditions for binding of DNA to the column, 2 volumes of QBT (see QPKH) were added to the column before application of the DNA. Final elution of DNA from the column was with elution buffer pre-heated to 50°C as opposed to room temperature (as recommended, QPKH).

Plasmid, cosmid and BAC clones were purified using ion exchange columns, and were found to be of sufficient quality for direct sequencing, restriction analysis and the construction of sub-clone libraries (section 3.4.2).

2.2.8 High throughput plasmid preparation

High throughput plasmid preparation was carried out on a Biomech™ 2000 robot (Beckman Coulter) using the QIAprep™ plasmid extraction reagents and the recommended automated protocol for plasmid extraction (Qiagen). Plasmid extractions were carried out in 96 well format. Purified plasmids were of sufficient quality to sequence directly (section 2.6.5) or analyse by restriction digestion (section 2.5.5).

2.3 Preparation of nucleic acids from tissues

DNA and RNA were extracted primarily from human tissues although zebrafish and mouse tissues were also used. Due to the hazardous nature of chemicals used in all of these protocols and the potentially pyrogenic status of samples, all of the described procedures were carried out in level 2 bio-safety environments.

For procedures involving the extraction or manipulation of RNA, all solutions were either DEPC (Diethylpyrocarbonate) treated or made up with DEPC treated water under clean room conditions. DEPC treatment consisted of adding 1/1000th volume DEPC to the solution, shaking vigorously and standing over night. The DEPC was subsequently removed by autoclaving of the solution. This treatment chemically modifies histidine and tyrosine side chains, consequently eliminating the activity of proteins including RNases.

2.3.1 Total cellular RNA isolation

Total cellular RNA was isolated from solid tissue following the protcol provided with RNA-STAT-60 (BioGenesis). Solid tissue samples (100 to 200 mg), previously stored in liquid nitrogen or at -70°C were homogenised in 2 ml of RT-STAT-60. Homogenisation was carried out using a motorised homogeniser with disposable heads to minimise the risk of contamination and eliminate any cross-contamination of samples. Chloroform extraction and nucleic acid precipitation were carried out as recommended (BioGenesis RNA-STAT-60 data sheet). Recovered RNA was resuspended in 100 µl of DEPC treated TE (pH 7.5) and stored at -70°C.

2.3.2 Purification of leukocytes for nucleic acid isolation

This method was developed to overcome problems of the dilute cellular concentration of whole blood and enucleated nature of mammalian erythrocytes. This protocol was designed to be rapid and flexible as it allows RNA and / or DNA isolation from leukocytes. 10 ml whole blood in EDTA (to prevent coagulation) was mixed by inversion with 10 ml of red cell lysis buffer (7.7%(wv) ammonium chloride, 0.8% (w/v) sodium bicarbonate) (Miller *et al.,* 1988). The emulsion was incubated on ice for 10 minutes. The degree of light refraction decreased during this

time as erythrocytes, but not leukocytes were osmotically lysed. Leukocytes were collected by centrifugation at 4,000 g for 10 minutes. After the supernatent was decanted, pellets of leukocytes were resuspended in 1 ml of RNA-STAT-60 and nucleic acids extracted as described in sections 2.3.1 and 2.3.4.

2.3.3 DNA isolation from solid tissue

Approximately 200 mg of frozen tissue sample (-70°C or liquid nitrogen) was ground up to a fine powder in mortar and pestle. The powder was mixed into an emulsion with 850 ml of dH_2O, 100µl of proteinase K buffer (50 mM Tris.HCl, pH 8.0; 50 mM KCl; 10 mM EDTA; 1mM dithiothreitol; 0.5% (w/v) SDS; 100 mg/ml RNAse A) and 50 µl of proteinase K (Sigma) at 10 mg/ml. The emulsion was incubated at 37°C for 2 hours with occasional agitation. The emulsion was subsequently phenol-chloroform extracted (section 2.4.3), nucleic acids were ethanol precipitated (section 2.4.1) and DNA was resuspended in 100 µl TE (pH 7.5) and stored at 4°C. This protocol was based on that described by Westerfield, (1995).

2.3.4 RNA-DNA parallel isolation

RNA extraction was carried out as described in section 2.3.1. Supernatant from the chloroform extraction step of the RNA extraction protocol (section 2.3.1) was mixed by inversion with 800 µl of DNA-STAT-50 (BioGenesis). The subsequent emulsion was phenol – chloroform extracted (section 2.4.3). Isopropanol precipitation (section 2.4.2) was used to extract DNA from the phenol – chloroform supernatent. The recovered DNA was washed in 75% ethanol, air dried and resuspended in 400 µl TE (pH 7.5). Recovered DNA was stored at 4°C.

2.4 Purification of nucleic acids

2.4.1 Ethanol precipitation

DNA and RNA were precipitated from aqueous solution in the presence of 2 volumes of 100% ethanol. If the salt concentration of the solution was low, 0.1 volumes of 3 M sodium acetate was added to aid the precipitation. The precipitation of small quantities of nucleic acids was aided by incubation at low temperature (4°C

to -20°C) for up to eight hours. Precipitate was collected by centrifugation at 10,000 g for 30 minutes, typically at 4°C.

Ethanol precipitation also caused the precipitation of ionic salts, these were removed by washing the pellet in 70 or 75% ethanol at room temperature and briefly re-centrifuging. This salt removal step was repeated if a large quantity of salt was present or the nucleic acid was to be used in a particularly salt sensitive application. Pellets were allowed to air dry prior to resuspension in TE (pH 7.5) or dH_2O.

2.4.2 Isopropanol precipitation

An alternative method to ethanol precipitation, isopropanol precipitation was used to minimise salt co-precipitation with nucleic acids. Less isopropanol was required for precipitation than with ethanol, providing a logistical advantage over ethanol in some protocols. Typically, 0.6 volumes of propan-2-ol (isopropanol) at room temperature were added to the nucleic acid preparation, mixed by inversion and incubated at room temperature for 10 minutes. The precipitate was pelleted by centrifugation and washed in 70 or 75% ethanol to remove salt residues, and replace isopropanol with the more volatile ethanol, to aid drying and resuspension. Isopropanol pellets were glassy in appearance and often difficult to see.

2.4.3 Phenol – chloroform extraction

Alkaline lysis supernatant was initially precipitated with 0.6 volumes of propan-2-ol, washed in 75% ethanol and resuspended in 200 to 750 µl of TE (pH 7.5). This initial precipitation was not necessary when phenol-chloroform extraction was being carried out in other contexts. 0.5 volumes of phenol and 0.5 volumes of chloroform were added and mixed by inversion. The emulsion was allowed to stand for 10 minutes at room temperature, then centrifuged at 10,000 g for 30 minutes. Nucleic acids remained in the upper aqueous phase, while denatured protein was precipitated at the interphase. Supernatant was decanted, then 1 volume of chloroform was added and mixed by inversion. The chloroform only step was included to remove residual phenol contamination, a potent inhibitor of many enzymatic reactions. After centrifugation, the aqueous top layer was decanted and precipitated with 0.6

volumes of propan-2-ol, the pellet washed in 75% ethanol and resuspended overnight at 4°C in TE (pH 7.5).

2.4.4 Recovery of DNA from agarose gels

DNA was recovered from low melting point (LMP) agarose (Gibco BRL) by one of two methods: spin-column purification or agarase digestion. Spin column purification was faster and recovery more efficient, but DNA molecules in excess of 10 kbp were sheared during the procedure. Agarase digestion was therefore the method used for large DNA molecules. For each of the following methods, DNA was electrophoretically separated (section 2.7.2) in a low melting temperature agarose gel made up with 1x TAE buffer (section 2.7.1). The target band was excised using a clean scalpel blade while briefly visualised by short wave, UV transillumination.

Spin column purification

The QIAquick gel extraction kit (Qiagen) was used according to manufacturer guidelines for rapid extraction of DNA from LMP agarose gel. DNA was typically eluted in 30 µl of dH_2O rather than 50 µl TE (the manufacturer recommended resuspension medium), to give a higher concentration and overcome problems of EDTA inhibition of some enzyme reactions. For DNA fragments in excess of 1 kb, the elution liquid was preheated to 55°C to improve the efficiency of recovery.

Agarase digestion

Agarase (Roche) enzymatically digests agarose into oligosaccharide fragments, DNA was then recovered by ethanol precipitation. The gel slice was diluted with 1× TAE so that the agarose concentration was not more than 0.8% (w/v) and 0.04 volumes of 25× agarase buffer (750 mM Bis-Tris, 250 mM EDTA) was added. Agarose was melted by heating to 70°C for 10 minutes and subsequently cooled to 37°C, then 1 U of agarase was added per 100 µl of solution. The reaction was allowed to proceed for 4 hours with occasional agitation. DNA was recovered by phenol-chloroform extraction (section 2.4.3) and ethanol precipitation (section 2.4.1).

2.4.5 Drop dialysis

The salt content of aqueous solutions was manipulated by placing individual drops (not more than 5 µl per drop) of solution on filter discs of 0.025 µm pore size (Millipore). Filter discs were floated on a pool of buffer / dH_2O of the desired concentration. Osmotic equilibration between the excess of solution below the disc and the small volume of liquid on the disc effectively adjusts the solution on the disc to that below the disc. The small pore size and properties of the disc polymer prohibit exchange of macromolecules, providing a rapid means of desalting nucleic acid preparations with minimal loss of nucleic acid. Drop dialysis was carried out for 30 minutes to allow equilibration. The sample drops were then recovered from the disc by pipetting.

2.5 Manipulation of nucleic acids

2.5.1 DNA ligation

DNA ligation is the covalent joining of DNA molecules through the 5' phosphate of one molecule and the 3' –OH of the other molecule. Ligation was used in the construction of clone libraries, by the insertion of target molecules into a linearised vector. Through 5' phosphorylation or dephosphorylation (sections 2.5.3 and 2.5.2) of DNA molecules, ligation was directed to prevent re-ligation of vector DNA without insert or to give a ligation reaction directionality. The ratio of substrates was also modulated to bias the production of one potential product over another, for example, substantially more insert than vector led to concatamerisation of the intended insert, rather than the majority of inserts being one target in one vector.

T4 DNA ligase (Roche) was incubated in 1× ligation buffer (66 mM Tris.HCl, pH 7.5; 5mM $MgCl_2$; 1 mM DTT) and 1 mM ATP, with the DNA substrates. Incubation was carried out at between 4°C and 20°C (typically 16°C) for at least 8 hours. Addition of polyethyleneglycol (PEG) to the reaction substantially enhanced the rate of ligation through macromolecular aggregation, used when following the protocol of Boyd, (1993). If DNA ligation was being used to clone a target molecule

into a vector, the ligase was heat killed at 65°C for 30 minutes and the reaction mix drop dialysed (section 2.4.5) over dH_2O prior to transformation (section 2.5.9).

2.5.2 Mono – 3' – adenylation of dsDNA

Mono-3'-adenylation adds a single adenosine base overhang onto the 3' end of a blunt ended, double stranded DNA molecule. A reaction was set up with 50 to 250 ng DNA template, 0.6 mM $MgCl_2$, 1 mM Tris.HCl (pH 8.3), 5 mM KCl, 0.2 mM dATP and 0.5 U Taq DNA polymerase. The reaction was incubated at 75°C for 15 minutes. The results of the products of the reaction were used directly in TOPO and T/A cloning (section 2.5.3). This protocol was devised through logical modification of the PCR protocol (section 2.6.2). It was typically used to increase the efficiency of TOPO and T/A cloning of PCR products. It was found to be particularly useful if the PCR product had been isolated and purified (section 2.4.4), and essential if a polymerase with proof reading activity was used for PCR.

2.5.3 TOPO cloning

As an alternative to the DNA ligase method of DNA ligation, the enzyme TOPO isomerase was used. Vector prepared with TOPO isomerase covalently bound to the ends was commercially available (Invitrogen). The Invitrogen TOPO cloning kit (Version E2) was used for the rapid and efficient cloning of PCR products. If the cloning target was a blunt ended restriction fragment, from high fidelity PCR (proof reading) or the PCR product had been purified (section 2.4.4) then 3' monoadenylation (section 2.5.2) was carried out prior to TOPO cloning. The TOPO cloning protocol was carried out as described in the manufacturer's instructions, using OneShot™ chemically competent cells and either pCR2.1 or pCR4.0 vectors (section 2.2.4). Transformants were selected on the basis of growth on selective agar plates and blue / white colour screening (section 2.2.3).

2.5.4 Electroporation

Electro-competent cells (XL1-Blue) were prepared by the protocol described in Sambrook et al (1989). Competent cells stored at -70°C were defrosted on ice, 2 to 8 µl of drop-dialysed (section 2.4.5) ligation mix was added to cells and mixed gently. Cells were decanted to an electroporation cuvette (10 mm separation

between electrodes) and shocked at 2.5 kV and 200 Ω for < 0.1 seconds using a Gene Pulser (Biorad). Shocked cells were mixed with 100 µl of pre-warmed LB and incubated for 55 minutes at 37°C in a shaking incubator. Cells were then spread onto appropriately selective agar plates at 1:10, 1:100 and 1:1000 dilutions and grown overnight at 37°C on appropriately selective agar plates (section 2.2.3).

2.5.5 Restriction digestion

DNA was diluted to the desired concentration, and hydrolysed at specific sites by type II restriction endonucleases in the presence of an appropriate buffer (as supplied by the restriction enzyme manufacturer, Roche). Type II restriction enzymes hydrolyse double stranded DNA at (or near) enzyme specific palindromic sequences to leave 5' overhang or 3' overhang sticky ends or blunt ends to the DNA.

Where enzymes worked effectively under the same salt conditions and multiple digests were required, enzymes were combined in the same reaction. If multiple enzymes were required but each worked optimally in incompatible salt conditions, restriction was carried out with one set of compatible enzymes, the reaction drop dialysed (section 2.4.5), the appropriate buffer and next set of compatible enzymes added and incubated. The salt and temperature conditions required by each enzyme were as recommended by the manufacturer (Roche).

Typical reactions contained a final DNA concentration of not more than 0.25 µg/µl. Restriction enzymes were used at 2 enzyme units per 1 µg DNA and incubated for more than 1 hour at the optimal temperature for enzyme activity (Roche). One enzyme unit can hydrolyse 1 µg of DNA in 1 hour under optimal conditions. The one fold excess of enzyme was used to ensure complete digestion of the DNA.

2.5.6 5' dephosphorylation

The removal of 5' phosphate groups from a DNA molecule prevents circularisation of that molecule during ligation reactions, particularly preventing recircularisation of empty vector during cloning. 5µg of vector DNA was restriction digested (section

2.5.5), with appropriate endonuclease(s), the reaction was drop dialysed (section 2.4.5) over dH_2O and subsequently incubated with 1 U calf intestinal phosphatase (CIP, Roche) in 1× CIP buffer (10 mM Tris.HCl, pH 8.3; 1 mM $ZnCl_2$; 1 mM $MgCl_2$). The reaction was allowed to proceed with occasional agitation for 30 minutes, then stopped by addition of 0.2 volumes of 0.5 M EDTA. The reaction product was size selected and purified as described in section 2.4.4.

2.5.7 5' phosphorylation

5' phosphorylation of DNA was carried out both to mediate cloning of PCR products into dephosphorylated vector (section 2.5.6) and the end-labelling of DNA molecules. Polynucleotide kinase (PNK, Roche) transfers the γ-phosphate of ATP to the 5' carbon of single or double stranded nucleic acids.

In a total volume of 50 µl, approximately 5 µg of blunt ended, 5' overhang or single stranded DNA was incubated in 1× kinase buffer (5 mM Tris.HCl, pH 7.5; 1 mM MgCl2; 0.5 mM DTT) with 10 U of PNK at 37°C for 30 minutes. PNK activity was removed by phenol-chloroform extraction (section 2.4.3) and DNA recovered by ethanol extraction (section 2.4.1). 5' phosphorylation of oligonucleotides was carried out during oligonucleotide synthesis (section 2.6.1) or labelling (section 2.8.1).

2.5.8 DNA digestion

Total hydrolysis of DNA was achieved with DNase I digestion under appropriate salt and temperature conditions. This method was used to remove DNA contamination from preparations of other molecules, typically RNA. 10 U of DNase I (Roche) was used per 1 µg of anticipated DNA contamination. Following digestion, DNA digestion was carried out in 1× DNase buffer (25 mM Tris.HCl; 50% (v/v), pH 7.6), for 15 minutes at 37 °C. DNase I was heat inactivated at 65°C for 30 minutes and / or removed by phenol-chloroform extraction (section 2.4.3).

2.5.9 RNA digestion

RNase A was used extensively for the removal of RNA during the preparation of DNA from cells and tissue. RNase A is highly thermal stable and enzymatically

active over a wide range of salt conditions and temperatures. It was therefore added directly to preparations without the addition of specific buffers. DNase activity was removed prior to use by heating the RNase A solution to 98°C for 30 minutes and cooling slowly to room temperature. RNase A was typically used at a working concentration of 0.1 µg/µl and incubated at 37°C for 15 minutes. Where necessary, RNase A was removed using 2 consecutive phenol-chloroform extractions (section 2.4.3).

2.6 Synthesis of nucleic acids

2.6.1 Oligonucleotide design and synthesis

Oligonucleotides for PCR (section 2.6.2) were designed using the Primer3 program (section 2.11.2), and screening against mis-priming libraries for human or mouse as appropriate for their final use. Typically oligos for PCR were 18 to 24 nucleotides long with approximately 50% G+C content and predicted to form minimal or no internal hairpin structures. Primer pairs for PCR were screened against possible primer dimer formation and were designed to have melting temperatures (Tm) within 3°C of each other.

Sequence walking primers were designed using consed (section 2.11.2) with an optimum length of 18 nucleotides and predicted Tm of 55 to 59°C. Where primer picking programs were unable to select appropriate oligonucleotides, they were designed manually, minimising hairpin structures, especially at the 3' end, and minimising any potential mispriming events. Primer candidates were also screened against a mis-priming library (using the lalign component of the Fasta2 package, section 2.11.2).

Oligonucelotides were commercially synthesised (MWG Biotech or Sigma Genosys), typically at the 0.01 µmol scale with standard purification Oligonucleotides in excess of 45 bases were synthesised at the 0.05 µmol scale and HPLC (High Performance Liquid Chromatography) purified.

2.6.2 Polymerase chain reaction

The polymerase chain reaction (PCR) is a cornerstone of molecular biology, allowing the rapid and specific amplification of a DNA template from a (potentially) highly complex mixture of DNA. Many optimisations and modifications to the basic technique have been used during this work to achieve specific goals. Typically a master mix of 1× PCR buffer (1 mM Tris.HCl (pH 8.3); 5 mM KCl), 1.0 to 2.0 mM $MgCl_2$, 1 ng/µl of each dNTP and 1 U/µl Taq DNA polymerase (PE Biosystems); was prepared on ice. Oligonucleotide primers at a final concentration of ~0.3 µM and template DNA (~1 ng for genomic DNA, 0.1 ng for plasmid DNA) were mixed on ice in reaction tubes. Master mix was aliquoted into reaction tubes and mixed with primer-DNA solution.

For specific applications such as high fidelity PCR or long range PCR, Taq polymerase was replaced by other polymerases and polymerase mixes. Where alternate polymerases were used, it has been documented in the results chapters. Buffer conditions for non-Taq polymerases were optimised using manufacturer recommendations.

The PCR thermal profile consists of an initial denaturing step, normally 94°C for 4 minutes, followed by 20 to 35 cycles of annealing, polymerase-mediated extension of primer, and denaturing. Annealing temperature was determined by the predicted Tm (melting temperature) of the primers, with the annealing temperature being typically 2 -3°C lower than the lower Tm. 72°C was the optimal extension temperature for Taq polymerase. The denaturation step was usually 92°C for 20 seconds, although both temperature and time were increased for high GC content amplicons.

Annealing was typically carried out for 20 seconds and extension for 30 seconds, for amplicons under 1 kb. For longer amplicons, the extension time was increased up to a maximum of 10 minutes per cycle. Reactions were carried out on MJR-PTC 220 or MJR-PTC 200 thermal cyclers (MJ Research) using heated lids to minimise

evaporation and "Sim-tube" thermal control. All reactions were carried out in thin-walled 0.5 ml or 0.2 ml tubes, or 0.2 ml 96 well plates.

2.6.3 First strand cDNA synthesis

First strand cDNA synthesis is the reverse transcription of an RNA molecule into a complementary DNA (cDNA) sequence that can be used as the substrate for "cloning" (after the synthesis of a second strand) and polymerase chain reaction (section 2.6.2). First strand cDNA synthesis was carried out using the Roche™ 1st strand cDNA synthesis kit. The cDNA synthesis was carried out as described in the cDNA synthesis kit data sheet. RNA templates were DNase treated (section 2.5.8) prior to 1st strand synthesis. Approximately 1 µg of total cellular RNA was used as the template for 1st strand synthesis reactions. For every 1st strand synthesis reaction, a second reaction was carried out that lacked the reverse transcriptase enzyme (AMV reverse transcriptase) for use as a negative RT-PCR control. Either random hexameric oligonucleotides or oligo dT oligonucleotides were used to prime first strand synthesis (indicated in the results chapters). Random hexamer primers were used at a concentration of 1 ng/µl and oligo dT at 0.5 ng/µl. The products of 1st strand synthesis were stored at -70°C.

2.6.4 Dideoxy sequencing

Rhodamine dye terminator master mix (Perkin Elmer) was initially used. BigDye chemistry terminators (Perkin Elmer) were subsequently shown to give improved signal consistency and longer read lengths. When plasmid templates were being used, 50% of the BigDye master mix was substituted for 4 µl of Halfterm buffer (200 mM Tris.HCl; 5 mM $MgCl_2$), economising on reagent costs with negligible effect on trace quality or length. For PCR product or plasmid templates, reads of 600 to 800 bp were typical. For cosmid templates this was reduced to 400 - 600 bp.

2.7 Electrophoresis

2.7.1 Solutions

TBE (×20)

1 M Tris.HCl, pH 8.0; 20 mM EDTA; 1M boric acid, pH 8.3.

TAE (×20)

0.8 M Tris.HCl, pH 8.0; 20 mM EDTA; 0.4 M acetic acid.

DNA loading buffer (×10)

20% (w/v) ficoll; 100 mM EDTA; orange G (Sigma).

RNA loading buffer (×10)

1 mM EDTA, pH 8.0; 0.25% (w/v) bromophenol blue; 0.25% (w/v) xylene cyanol; 50% (v/v) glycerol.

MOPS running buffer (×10)

0.4 M MOPS (3-(N-morpholino)-propanesulfonic acid), pH 7.0; 0.5 M sodium acetate; 0.01 M EDTA.

SSC (×20)

3 M NaCl; 0.3 M $Na_3C_6H_5O_7$ pH 7.0

2.7.2 Agarose gel electrophoresis

Size fractionation of nucleic acids was achieved by migration through an agarose matrix in an electric field. Agarose MP (Roche) was dissolved in 1× TBE or 1× TAE buffer by heating. The final concentration of agarose was within the 0.4% to 4% (w/v) range dependent on the size of nucleic acid fragments to be efficiently resolved. RNA was typically separated under denaturing conditions (section 2.7.3). Digested genomic, BAC or cosmid DNA was typically run on 0.6 to 0.8% gels; plasmid restriction enzyme digests and PCR products were run on 1% to 2% gels.

DNA samples were loaded into wells with $1/10^{th}$ volume DNA loading buffer. The gel was placed in an electrophoresis tank of appropriate size and sufficient 1× TAE or TBE (to match the gel) buffer was added to just submerge the gel. A uniform, direct electrical current at 25 to 100 V was passed through the buffer, causing nucleic acids and dye in the loading buffer to migrate to the anode. Nucleic acid fragments were visualised by including the dye ethidium bromide in the gel and running buffer at a final concentration of 0.1 μg/ml. The dye intercalates between the bases of nucleic acids, when intercalated ethidium bromide fluoresces in the visible spectrum upon stimulation by UV translillumination (optimal λ 260 nm).

DNA size markers

Standardized DNA size markers were run alongside samples during electrophoresis, to allow size estimation of DNA bands. The marker used depended on the expected range of band sizes to be resolved. For most PCR products and plasmid restriction analysis, the ΦX174 *Hae III* (Roche) or 100 bp DNA ladder (Promega) were used. Where larger DNA fragments were to be resolved, the 1 kb DNA ladder (Gibco BRL) and λ *Hind III* restriction fragments (Roche) were used as markers. Typically 200 ng of DNA was loaded per marker lane, and at least 2 marker lanes were run per gel to check for even migration across the gel.

2.7.3 Denaturing agarose gel electrophroesis

Extensive secondary structure formed by RNA under non-denaturing conditions disrupts the correlation of linear size and electrophoretic mobility. Consequently, RNA was size separated under denaturing conditions. RNA was loaded into a gel of appropriate agarose concentration (section 2.7.2), 1× MOPS running buffer and 2.2 M formaldehyde. The gel was submerged in 1× MOPS running buffer during electrophoresis. Formaldehyde concentration of the gel was increased if the run time was expected to exceed 3 hours. As formaldehyde concentration increases, gels become more brittle.

0.5 to 10 μg of RNA in 11 μl TE was mixed with 5 μl 10× MOPS running buffer, 9 μl 12.3 M formaldehyde and 25 μl formamide, incubated for 15 minutes at 55°C and

mixed with 10 µl RNA loading buffer before loading into wells of the denaturing gel. Gels were run at 100 V until the required degree of migration was achieved. After running, RNA could be visualised by UV transillumination after soaking the gel in 0.5 M ammonium acetate and 0.5 µg/ml ethidium bromide for 50 minutes. Alternatively, ethidium bromide was mixed with the RNA to a final concentration of 0.5 µg/ml prior to loading into the well.

2.8 Labelling of nucleic acids

2.8.1 Oligonucleotide labelling

Oligonucleotides were labelled using polynucleotide kinase (Roche) as described for phosophorylation of DNA (section 2.5.2). For radiolabelling, [γ-^{32}P]ATP was substituted for ATP as the phosphate donor. Unincorporated radioactivity was removed by gravity flow through a Sephadex® G25 column (Amersham Pharmacia Biotech).

2.8.2 Labelling of dsDNA

Double stranded DNA was labelled using the random primed method (Feinberg and Vogelstein, 1983; Feinberg and Vogelstein, 1984). The High Prime (Roche) reagent and protocol was used with [α-^{32}P]dCTP (800 Ci/mmol). Unincorporated nucleotides were removed by gravity flow through a Sephadex® G50 column (Pharmacia Biotech). Probe DNA was denatured by heating to 99°C for 10 minutes in the presence of 50 µl unlabelled, sonicated salmon sperm DNA (10 mg/ml) prior to adding it to the hybridisation (section 2.10).

2.8.2 Riboprobe synthesis

Riboprobe template constructs were generated by PCR (section 2.6.2) using one standard primer and one primer incorporating the 19 bp T7 promoter sequence (5'-TAATACGACTCACTATAGG-3') at the 5' end of the oligo. Where possible, target sequences for amplification were chosen so that the limiting nucleotide (cytosine) was not incorporated into the transcript within 12 nucleotides of the transcriptional start, thereby minimising premature transcript termination (Ling *et al.*, 1989). Riboprobe templates were PCR amplified from plasmid clones or a

parent PCR product that had been size selected and purified using the protocol described in section 2.4.4.

The MAXIscript™ T7 transcription kit (Ambion) reagents and protocol were used for transcription of the template. Template DNA was typically used directly from the PCR reaction without further purification. In problematic cases, the template was purified by the method described in section 2.4.4. The labelled nucleotide was [α-^{32}P]CTP (800 Ci/mmol) with 50 µCi used per reaction. Labelled CTP was the only source of CTP in the transcription reaction, resulting in probes of a high specific activity. Because of the high specific activity and generally unstable nature of purified RNA, riboprobes were used for hybridisation on the same day as synthesis.

Prior to use in hybridisation (sections 2.10.4 and 2.10.5), riboprobes were purified from unincorporated nucleotides by gravity flow through a Sephadex® G50 column (Pharmacia Biotech).

2.9 Membrane immobilised nucleic acids

2.9.1 Solutions

SSC

3 M NaCl; 0.3 M $Na_3C_6H_5O_7$ pH 7.0

Denaturing solution

0.5 M NaOH; 1.5 M NaCl

Neutralising solution

1 M Tris.HCl; 2 M NaCl; pH adjusted to 5.5 with HCl.

2.9.2 Southern transfer of DNA

DNA was transferred from agarose gel to nylon membrane by capillary blotting, adapted from Southern (1975). After electrophoresis of DNA, the gel was

photographed while transilluminated with UV light at λ260 nm, next to a ruler for scale.

Double stranded DNA with an expected size of >500 bp was denatured with denaturing solution (section 2.9.1) and subsequently soaked in neutralizing solution (section 2.9.1) prior to capillary transfer. Denaturing and neutralizing was for an equal period of time, dependent upon the length of DNA molecules to be transferred and agarose concentration of the gel. Longer denaturation was carried out for larger DNA molecules and higher agarose concentrations, up to a maximum of 60 minutes. Gels were subsequently washed in 2× SSC (section 2.9.1) for 10 minutes, a step that was found to substantially reduce subsequent hybridisation background signal.

The capillary transfer was set up in the following manner: 5 mm filter paper (Whatman) was soaked in 20× SSC and draped across a glass plate with its ends dipped in a reservoir of 20× SSC. A further sheet of 5 mm paper, larger than the gel, was also soaked in 20× SSC and placed on the first. The gel was placed directly on the top sheet of filter paper. All exposed filter paper was sealed with Saran Wrap (Dow Chemical Company). A nylon membrane (Osmonics) cut to the same width and length of the gel was soaked in 2× SSC and placed directly onto the gel. A sheet of 3 mm filter paper, was soaked in 2× SSC and placed on the membrane, then a further 2 sheets of 3 mm filter paper were added onto the stack. Air bubbles were carefully excluded at each stage of layering. At least 8 cm depth of paper towels was stacked and weighted down below a glass plate directly on top of the gel-filter paper stack. Capillary transfer was allowed to proceed for a minimum of 8 hours.

Following capillary transfer, the position of the gel wells could be clearly seen and was marked on the membrane with a pencil. The membrane was allowed to air dry, while being protected from contamination by placing it between dry 3 mm filter paper. DNA was then covalently bound to the membrane with 1200 $\mu J/cm^2$ of UV irradiation. Membranes were stored at room temperature until use, protected by filter paper.

2.9.3 Northern transfer of RNA

The protocol for transfer of RNA from agarose gel to a nylon membrane was the same as for DNA (section 2.9.1), however 10× SSC rather than 20× SSC was used for capillary transfer. RNA was typically separated on a denaturing gel negating the need for further denaturation prior to capillary transfer. Prolonged exposure to UV transillumination was sufficient to cleave large RNA molecules aiding their efficient capillary transfer. All solutions were DEPC treated (section 2.3) or made up with DEPC-treated water. Lab-ware used during Northern transfer was washed in RNase-Zap® (Ambion) prior to use.

2.9.4 Spot-blotting

Spot-blotting was a rapid method of binding DNA molecules to a nylon membrane for subsequent hybridisation. Approximately 1 cm diameter circles were drawn on a nylon membrane (Osmonics) using a Papermate 2000™ ball point pen. 2 µl drops of DNA solution at 0.25 to 1.0 µg/µl concentration were spotted in the centre of each circle. The spot was allowed to air dry. Subsequent spots were added and air dried until the final desired DNA concentration was reached (typically 1 µg). The membrane was then baked at 80°C for 60 minutes to securely bind the DNA to the membrane. Filter paper (3 mm Wattman paper) was soaked in denaturing solution, a second filter paper soaked in neutralising solution and a third in 2× SSC. The nylon membrane was then sequentially placed on the denaturing, neutralising and finally 2× SSC soaked filter papers, for 3 minutes each. The membrane was then air dried and DNA covalently bound to the filter by UV radiation (1200 µJ/cm^2).

2.9.3 Preparation of library filters

Sub-clone library filters were prepared using a Flexsys robot (PBA Technology) to spot individual clones from 96 well plates onto a 7 × 11 cm nylon membrane (Osmonics). Each clone was double spotted into a 3 × 3 array, with 96 such arrays on the filter (figure 2.1). Gridded filters were laid on 7 × 11 cm agar plates, avoiding air bubbles. Colonies were grown for 16 hours at 37°C and subsequently processed as described in section 2.9.5.

a b c

c d

b d a

Figure 2.1; Library filter gridding pattern. Each letter a to d represents a different 96 well plate. The spotting pattern was designed to allow rapid and unambiguous assignment of plate number, and the grid coordinate of the 3 × 3 array corresponded to the same coordinate in the master plate. Spotting was carried out to PBA Technology recommendations.

2.9.4 Bacterial colony lifting

Bacterial colonies grown on solid media were transferred to a Protran BA 85/20 nitrocellulose disc (Schleichler and Schulell™), by gently laying the disc on the surface of the agar plate. An asymmetric pattern of marks on the disc was copied to the petri dish while the disc was in place, thereby allowing future alignment of an autoradiograph and the plate. The disc was lifted from the plate after 2 minutes; after this time the majority of cells from individual bacterial colonies are attached to the disc. By lysing the bacteria and fixing cell lysate to the nitrocellulose disc (section 2.9.6) a pool of bacterial transformants were readily screened by hybridisation for possibly rare target inserts.

2.9.5 Bacterial on-filter lysis and fixation

Bacterial colonies bound to nitrocellulose membranes (section 2.9.4) were lysed on the filter and the cell lysate subsequently bound to the filter. The nitrocellulose membranes were sequentially layered on 3 mm Wattman filter paper that had been previously soaked in **(1)** 20% (w/v) SDS **(2)** denaturing solution, **(3)** neutralising solution, **(4)** 2× SSC. The membranes were layered on the pre-soaked filter papers for 1 minute, 3 minutes, 3 minutes and 2 minutes respectively. The membranes were allowed to air dry and were then heated at 80°C for 40 minutes under a vacuum. The 20% SDS lyses the bacterial cells. Sequential denaturing and neutralising resulted in single stranded DNA that was then bound to the membrane by heating to 80°C. Note, nitrocellulose is explosive and should only be heated in the absence of oxygen.

2.9.6 Libraries and commercial filters

In addition to the clone libraries generated during this work, several commercial or publicly available libraries were used as were commercial Northern blots.

Fugu genomic (cosmid)

A genomic DNA library from the puffer fish *Fugu rubripes*, made publicly available by the UK HGMP-RC in a gridded 4×4 array on four 22 × 22 cm Hybond N and one 7.8 × 11.9 cm Hybond N+ nylon membrane, was screened. Each of the 74,880 clones were spotted in duplicate. The library was made by Elgar and Nizetic (unpublished). The average insert size was approximately 40 kb representing an 8× genome coverage. The library was constructed in the Lawrist 4 cosmid vector (section 2.2.4). For further information on the library construction and gridding, see http://www.hgmp.mrc.ac.uk/ISO9000/BIOLOGY/LIBRARIES/fugu/fugu_filter_info.shtml.

Fugu genomic (BAC)

A genomic DNA library from the puffer fish *Fugu rubripes*. Produced by Incyte Genomics (library reference number: BAC-6253). Gridded double spotted filters for this library were available through a collaboration with G. Elgar (HGMP-RC, Hinxton, UK).

Fugu cDNA

10 *Fugu rubripes* cDNA libraries corresponding to tissues: "whole body", spleen, gill, gut, gonad, brain, eye, liver, kidney and "mixed tissue" (libraries A to J respectively) were obtained from the UK HGMP-RC. The libraries were constructed by Elgar, Warner and Hills (unpublished). Each library contains 36,864 clones double spotted in 4 × 4 configuration, on one 22 × 22 cm Hybond N nylon membrane. The libraries were constructed in *EcoRI/XhoI* cut pBluescript II KS+ plasmid (Stratagene), and the inserts were directionally cloned.

Mouse genomic RPCI-23 (BAC)

A mouse genomic (BAC) library produced from a female C57BL/6J mouse. The standard reference library for the UK Mouse Sequencing Consortium[1] and the BAC

fingerprint map of the mouse genome[2] . The RPCI-23 library was constructed by Osoegawa *et al.,* (2000). The cloning vector back bone for the library is the pBACe3.6 vector that includes the chloramphenicol resistance gene as a positive selective marker.

1. http://mrcseq.har.mrc.ac.uk/
2. http://www.gcgsc.bc.ca/projects/mouse_mapping/

Multiple tissue northern

Multiple tissue northern (MTN) blots were obtained from Clontech. The blots were prepared with poly-A selected RNA bound to a Hybond N+ membrane. MTN membranes were stored at room temperature until use, and subsequently at -70°C wrapped in Saran Wrap (Dow Chemical Company) after the first hybridisation. Hybridisation of MTN blots was carried out as described in sections 2.10.3 and 2.10.5. MTN blots could be successfully re-probed up to 3 times without significant loss of signal.

2.10 Hybridisation protocols

2.10.1 Solutions

Dextran hybridisation solution (DHS)

5× SSC; 10% (w/v) dextran sulphate; 0.1% sodium pyrophosphate; 0.1% SDS; 5× Denhardt's solution (0.1% (w/v) ficoll type 400; 0.1% (w/v) polyvinyl pyrrolidone; 0.1% (w/v) BSA (Sigma)).

Oligonucleotide hybridisation solution

5× SSC; 0.1% sodium pyrophosphate; 0.1% SDS; 5× Denhardt's solution (0.1% (w/v) ficoll type 400; 0.1% (w/v) polyvinyl pyrrolidone; 0.1% (w/v) BSA (Sigma)).

Church – Gilbert hybridisation solution

1 mM EDTA; 0.25 M $NaHPO_4$, 1% SDS.

Church – Gilbert wash solution

1 mM EDTA; 40 mM NaHPO4; 1% SDS.

Formamide hybridisation solution (FHS)

5× SSC; 0.1% SDS; 5× Denhardt's solution; 50% (v/v) formamide.

RNA : RNA hybridisation solution (RRHS)

5× SSPE; 0.1% SDS; 50% formamide.

2.10.2 DNA : DNA hybridisation

DNA bound to nylon or nitrocellulose membranes was washed in 2× SSC and placed between two sheets of nylon gauze. The gauze – membrane sandwich was rolled together ensuring the removal of air bubbles and placed in a glass hybridisation bottle (Hybaid) 1/3rd filled with 2× SSC. The bottle was rotated so that the membrane and flanking gauze unrolled against the surface of the bottle with air bubbles excluded.

The SSC was replaced by dextran hybridisation solution (DHS), the bottle sealed and incubated at the hybridisation temperature for 30 minutes. DHS was poured from the hybridisation bottle and replaced with a fresh aliquot of DHS, sufficient for a depth of 2 cm when the bottle was loaded in a hybridisation oven (horizontal). Pre-hybridisation was carried out at the eventual hybridisation temperature, typically 68°C for high stringency. Reduced stringency hybridisations between human and *Fugu* genomic sequence were carried out at 48°C.

Where it was useful to minimise background signal, sonicated salmon sperm (SSS) DNA was heat denatured and added to the pre-hybridisation solution to a final concentration of 10 µg/ml. For gridded membranes, a low level of background signal was often useful in determining accurate coordinates of positive signals. Labelled probe was added directly to the pre-hybridisation after denaturation (section 2.8). Hybridisation was typically carried out for 12 to 16 hours.

An alternative protocol was set up in the same manner, but used Church – Gilbert hybridisation solution. Church – Gilbert (pre-)hybridisations (*Fugu* BAC library screen, section 3.4.3) were carried out at 65°C.

2.10.3 DNA : RNA hybridisation

DNA : RNA hybridisations were set up in the same manner as described in section 2.10.2. All reagents were made up with DEPC treated ingredients; glass and plastic ware were washed in 5% (v/v) RNaseZap® (Ambion) prior to use.

2.10.4 RNA : DNA hybridisation

Riboprobe hybridisations were set up in the same manner as described for DNA : DNA hybridisation (section 2.10.2), with DEPC treated reagents and 5% (v/v) RNaseZap®-washed glass and plastic ware. All (pre-)hybridisation was carried out in formamide hybridisation solution (FHS). Hybridisation in FHS was carried out at 45°C for high stringency. The high background signal often observed with RNA probes was reduced by an extended pre-hybridisation of 4 hours and increasing the final concentration of SSS from 10 µg/ml to 100 µg/ml.

2.10.5 RNA : RNA hybridisation

The hybridisations were set up as described for DNA : DNA hybridisation. As with all other hybridisations involving RNA, solutions were made up with DEPC treated dH_2O, glass and plasticware was treated with RNaseZap. The High stability of RNA : RNA duplex was found to require more stringent hybridisation conditions than for DNA : RNA hybridisation, to prevent extensive non-specific annealing from taking place. RRHS was used with 5× Denhardt's solution and yeast total cellular RNA at 100 µg/ml for prehybridisation. The concentration of Denhardt's solution was reduced to 1× for hybridisation. Prehybridisation was carried out for 1 to 4 hours at 68°C, and hybridisation for 12 to 16 hours at 68°C.

2.10.6 Oligonucleotide hybridisation

Oligonucleotide hybridisation was set up as described for DNA : DNA hybridisation, oligonucleotide hybridisation buffer was substituted for DHS. The hybridisations were carried out at 10°C below the predicted (Primer3, section 2.11.2) primer melting temperature.

2.10.7 Washing conditions

Hybridised membranes were removed from hybridisation bottles and separated from nylon gauze. Membranes were then washed in 400 ml of 0.1× to 4× SSC, 0.1% SDS, with continual agitation. The temperature and salt concentration of the wash solution were dependent on the strength of the background signal and intended stringency of the hybridisation.

For high stringency DNA : DNA or DNA : RNA hybridisation, washing was carried out in 0.5× SSC at 70°C for 30 minutes to 3 hours dependent on the intensity of Geiger counter measured background signal. Washing of RNA probes under high stringency was carried out at 60°C in 0.5× SSC with 1% SDS for up to 2 hours. Oligonucleotide probes were washed at room temperature in 4× SSC and 0.1% SDS for up to 10 minutes. Church – Gilbert hybridisations were washed in Church – Gilbert wash solution at room temperature for 30 minutes, with constant agitation.

Reduced stringency conditions were used for cross-species hybridisation and identification of paralogous sequences. Reduced stringency was achieved by increasing the salt concentration of the wash solution (up to 4× SSC) and decreasing its temperature. Where reduced stringency conditions were used, they have been documented in the appropriate results sections.

2.10.8 Autoradiography

Membranes with bound, radioactively labelled probe were wrapped in Saran Wrap (Dow Chemical Company), avoiding creases and air bubbles. Wrapped membranes were taped into autoradiography cassettes and exposed to curix blue HC-S plus (AGFA) or BIOMAX™ MS (Kodak) film; the BIOMAX film was used for maximum sensitivity. Luminescent markers (Stratagene) were used for alignment. Exposure times depended on signal intensity, typically ranging from 10 minutes to 48 hours. For exposure times in excess of 1 hour, the autoradiography was allowed to proceed at -70°C, otherwise it was carried out at room temperature. Films were developed in a Fugi RG-II X-ray film processor.

Alternatively, the wrapped membranes were exposed to an erased phosphor screen (Molecular Dynamics). Signal from the phosphor screen was interpreted by a Storm II phosphorimager (Molecular Dynamics). The phosphorimager system converts the autoradiograph signal into a digital image, allowing quantification and comparison of signals more accurately than was possible with classic autoradiography and subsequent scanning of the developed film.

2.10.9 Membrane stripping and storage conditions

Membranes were stripped of hybridised probe by washing in strip solution (0.1× SSC; 0.1% SDS; 0.2 M Tris-HCl, pH 7.5) at 45°C for 30 minutes with constant agitation. If the membrane was a Northern blot (section 2.9.2), strip solution was made up with DEPC treated ingredients. For DNA bound membranes, any remaining signal was removed by washing in 0.1% SDS at 90°C for 30 minutes, with 2 changes of wash. This harsh treatment significantly reduced the number of subsequent probings possible; an alternative method was to allow residual signal to decay prior to reuse. Between uses, membranes were wrapped in Saran wrap to prevent drying and stored at -70°C.

2.11 Computational methods

The majority of software and databases were used unmodified. Appropriate versions and references are given and any modifications documented. Modifications to default parameters are indicated along with the associated results. Unless otherwise stated, all computational analysis was performed on Sun Ultra10 workstations or a Sun Enterprise 1000 server with 20 processors in a symmetric multiprocessing configuration and 5 Gb of random access memory. Both workstations and servers were based on the UltraSparc II processor and were running the Solaris 7 operating system.

2.11.1 Databases

Database	Notes	URL
Accession maps	Clone based assembly of the human genome	http://genome.wustl.edu/gsc/human/Mapping/
EMBL*	Nucleotide	http://www.ebi.ac.uk/embl/
ENSEMBL	Genome annotation project	http://www.ensembl.org/
Human FPC	Human BAC fingerprint contigs	http://genome.wustl.edu/human/
Interpro	Protein domain descriptions	http://www.ebi.ac.uk/interpro/
Mouse FPC	Mouse BAC fingerprint contigs	http://www.bcgsc.bc.ca/projects/mouse_mapping/
Mouse whole genome shotgun	Nucleotide (with quality data)	http://trace.ensembl.org/
OMIM	Human genes and inherited diseases	http://www3.ncbi.nlm.nih.gov/omim/
PFAM	Protein domains	http://www.sanger.ac.uk/Software/Pfam/
Prosite	Protein motifs	http://www.expasy.org/prosite/
PubMed	Literature	http://www.ncbi.nlm.nih.gov:80/entrez/
RepBase	Interspersed repetitive elements	http://www.grinst.org/
SPTR	Protein	http://www.ebi.ac.uk/swissprot/
Tetraodon whole genome shotgun	Nucleotide	http://www.genoscope.cns.fr/externe/tetraodon/
TIGR Microbial database	Finished and unfinished microbial genomes	http://www.tigr.org/tdb/mdb/
Transfac	Transcription factor binding sites	http://transfac.gbf.de/TRANSFAC/
UniGene	Nucleotide	http://www.ncbi.nlm.nih.gov/UniGene/
UTRdb	Database of UTR sequences	http://bighost.area.ba.cnr.it/BIG/UTRHome/
Zebrafish whole genome shotgun	Nucleotide (with quality data)	http://trace.ensembl.org/

Table 2.2; Summary of databases. * The EMBL database is subdivided into several groups including the EMBLminus data set that defines the high quality (relative to other biological sequences) annotated sub-set of sequences.

2.11.2 Software

Program/package	Version	Reference	Notes
Similarity searching and alignment			
AVID	NA	Bray *et al.*, In preparation	
BLAST	2.0.12	Altschul *et al.*, 1997	See comment **a**
BLASTZ	NA	Schwartz *et al.*, 2000	
ClustalW	1.74	Thompson *et al.*, 1994	
Cross_match	0.990329	Green, unpublished	
Dotter	3.0	Sonnhammer and Durbin, 1995.	
EMBOSS	2.0.1	http://www.uk.embnet.org/	
Est2Genome	NA	Mott, 1997	
FASTA2	2	Pearson *et al.*, 1997	
FASTA33	3.3	Pearson *et al.*, 1999	
NIX	NA	Williams, unpbulished	
Pipmaker (Blastz)	NA	Schwartz *et al.*, 2000	
Sim4	2001-02-06	Florea *et al.*, 1998	
SSAHA	2.0	Cox *et al.*, unpublished	
Wise2	2.1.20c	Birney, unpublished.	See comment **b**
Sequence assembly			
Consed	11.0	Gordon *et al.*, 1998	
Phrap	0.990329	Ewing and Green, unpublished	See comment **c**
Phred	0.990722.i	Ewing *et al.*, 1998	See comment **c**
***Ab initio* gene prediction**			
FGenes	1.6	Solovyev *et al.*, unpublished	
Genefinder	NA	Green, unpublished	
Genemark	2.2	McInich *et al.*, unpublished	
Genscan	NA	Burge and Karlin, 1997	
Grail2	2.0	Xu *et al.*, 1994	
HMMGene	1.1	Krogh, unpublished	
TSSW	NA	Solovyev, unpublished	
Protein sequence analysis			
Coils	NA	Lupas *et al.*, 1991	
Jpred	2	Cuff *et al.*, 1998	
Globe	NA	Rost, unpublished	

MITDISC	1.0	Nakai and Kanehisa, 1992	
Multicoil	1.0	Wolf *et al.*, 1997	
Phd	1996.1	Rost, 1996	
PIX	NA	Williams, unpublished	
PsortII	10.12.99	Nakai and Horton, 1999	
SEG	NA	Wootton and Federhen, 1996	
TMap	NA	Persson and Argos, 1994	
TMpred	NA	Hofmann and Stoffel, unpublished	
Sequence annotation and visualisation			
Alfresco	0.91	Jareborg and Durbin, 2000	
Laj	NA	Schwartz *et al.*, 2000	
NIX	NA	Williams, unpublished	
RasMol	2.5	Sayle and Milner-White, 1995	
RepeatMasker	7.7.2001	Smit, unpublished	
Vista	NA	Mayor *et al.*, 2000	See comment **d**
Phylogenetic analysis			
PHYLIP	3.57c	Felsenstein, unpublished.	
Mega2	2.1	Kumar *et al.*, 2001	See comment **e**
Sequence retrieval and data storage			
SRS	6.073	http://www.lionbio.co.uk/	See comment **f**
AceDB	4.8c	http://www.acedb.org/	

Table 2.3; Summary of software. See comments below.

Comments

(a) The single BLAST distribution (version 2.0.12) contained BLASTN, TBLASTN, TBLASTX, BLASTP, FASTACMD and FORMATDB functions. BLASTP could be used in Psi-BLAST mode. SEG compositional filtering of protein sequences and DUST filtering of nucleic acid sequences were also built into the same distribution and used by default.

(b) For the alignment of protein sequence to genomic sequence of more than 50 kb in length, the sequence coordinates of TBLASTN high scoring segment pairs were used to direct the alignment by the GeneWise component of

Wise2. When TBLASTN was used in this manner, a flat scoring matrix (20 points for a match, -5 for a mismatch) was used with gapping and compositional filtering switched off and the number of best hits from a region limited to one. This strategy substantially reduced the computational time required and reduced the number of erroneous gene structures when there was a high sequence identity between the protein and translated nucleic acid sequences.

(c) Phred and Phrap were used through the PhredPhrap Perl script (Ewing and Green, unpublished). For assemblies containing sequence fragments of more than 1000 nucleotides, the modified version of Phrap, Phrap.longrcads (distributed with Phrap) was used.

(d) Vista was typically used through the vistawrap interface, the source code for which is included in appendix II.

(e) The Mega2 software was run on a Pentium III machine running the Windows 98 operating system.

(f) The SRS database used was indexed, updated and maintained at the UK-HGMP-RC.

2.11.3 Manipulation of biological sequences

There are many software packages available to reformat sequences, isolate sub-sequences, mask regions of sequence and reverse complement sequence. However, these packages were primarily intended for interactive use rather than working in automated analysis pipelines dealing with data streams. To address these short comings in available software, a series of short Perl scripts (programs) were developed specifically to work efficiently with streams of sequence data. Raw Perl code for the principal programs: "grab", "revcomp" and "stfa" are included in appendix II. Grab isolates sub-sequences, masks (substitutes A, T, G or C nucleotides for N to prevent analysis of specific sequence regions) sub-sequences or masks all apart from specific sub-sequences as defined by sequence coordinates, in a stream of fasta format sequence data. The program revcomp simply returns the reverse complement of a nucleic acid sequence. Split fasta (stfa) recognises blocks of fasta format sequence in streams of data and writes them out into separate files.

Sequences of approximately one million base pairs or greater pose a particular problem when being manipulated. Typically a program will read an entire sequence into memory before attempting to manipulate it, for example, to obtain a sub-sequence. Reading large sequences into memory is in its self time consuming, if the machine does not have enough physical memory it will use the hard disc as virtual memory, slowing the process further. Neither the grab script described above nor other software generally available in molecular biology software packages provides a useful solution to this problem without pre-processing of the file to fragment it or index the content in some way.

A novel solution to this problem has been developed where the relative time taken to retrieve a sub-sequence is dependant on the length of the sub-sequence, not the parent sequence. This method depends on the assumption that every line of sequence in a fasta format file is the same length apart from the last line. This is a reasonable assumption if the file has not been manually edited. By reading the first two lines of a fasta file, the coordinates in bytes of a specific sub-sequence can be calculated. Using these byte coordinates a specific section of the file can be read directly from the disk without having to read the remainder of the sequence file. This method has been implemented as a Perl module (called sgrab) with an object oriented wrapper for its simple integration into other Perl scripts. Code for this module and an example of its use in a simple Perl script are detailed in appendix II. Table 2.4 details a simple benchmark comparison between several pieces of software performing an identical task on the same machine.

Program[a]	Package[b]	System time[c]	Elapsed time[d]
grab	NA	Failed	Failed
sgrab	NA	0.03	0.08
extract-fasta	MUMer	2.43	22.39
extractseq	EMBOSS	11.26	457.07
BioPerl	Bioperl	28.48	1631.77

Table 2.4; Benchmark testing of sub-sequence programs. **(a)** The program being compared. BioPerl refers to a simple Perl script wrapper around the Bio::Seq object that is

typically used in BioPerl (http://www.bioperl.org/) to extract a sub-sequence. **(b)** The name of the software package that the program is distributed as a component of (see section 2.11.2). Grab and sgrab were written for this work and are not yet distributed **(c)** Processor time taken to complete the function. **(d)** Real time taken (seconds) to write the sub-sequence corresponding to bases 30,000,000 to 30,100,000 from the 47,662,662 nucleotide sequence of chromosome 22 to a file (/dev/null). The machine used was a Sun Ultra10 with 128 Mb of RAM running the Solaris 7 operating system, the chromosome 22 sequence was located on an EIDE disk local to the machine carrying out the function.

2.11.4 Iterative sequence clustering and assembly

Within the public nucleic acid sequence databases, there are large datasets where individual sequences are generally of poor quality and are largely or completely redundant. This is particularly true of unassembled whole genome shotgun data, expressed sequence tags (ESTs), high throughput genomic sequence (HTG) and genomic survey sequence (GSS). Through clustering of related sequences, quality can be improved via the generation of a sequence consensus and redundancy is intrinsically eliminated. This principal of clustering forms the basis for single clone sequencing projects (section 3.5), UniGene assemblies of expressed sequences (Schuler, 1997) and whole genome sequencing projects (Adams *et al.,* 2000; Venter *et al.,* 2001).

Several approached were taken to cluster nucleic acid sequences, dependant on the nature of the sequence, its expected quality and the resources available for that data set. Perl and unix shell scripts were utilised to automate this process, but the differing data sources and basic assumptions for each assembly required substantial modification of the pipeline for each application. For this reason a generalised implementation of the pipeline could not be made available.

Regardless of the exact method and cut off values used, the underlying method was the same (specific parameters are defined subsequently). An initial search of a specific data set was undertaken with a protein or nucleic acid sequence. Matching sequences with alignment scores or identity values above a predefined cut off were obtained (the 'seed' dataset) and assembled using Phrap (section 2.11.2). Phrap

assemblies were then appropriately masked for repeats and searched back against the original data set. All high scoring matches were obtained, assembled, masked and searched again. This process was iterated until no new sequences were incorporated into the assembly, or a complete assembly over the region of interest was produced. The high stringency of Phrap assembly also permitted the multiplexing of separate assemblies when multiple transcripts or regions of a genome were assembled (see the EST clustering section below).

EST clustering

Organism specific EST databases were generated using SRS (section 2.11.2) and used for both an initial search to define 'seed' datasets and for the subsequent iterative cycles. The initial search used TBLASTN with a bit score cut off of 58 to define the seed dataset. This cut off was used as it was found to represent a boundary in the distribution of scores between the best matches and all other sequences. Iterative searching was carried out using BLASTN with a bit score cut off of 80, this was empirically determined to maximise extension of contigs and identification of homologous non-identical sequences while minimising matches to unrelated sequences. As an additional step in EST clustering, when a cluster was unable to be extended further, it was searched against the EMBLminus database (section 2.11.1). Matches from EMBLminus were manually screened to determine if they represented transcripts rather than genomic DNA and if they were annotated as from the expected organism. If both of those criteria were met, the mRNA sequence was included in the next round of assembly.

Mouse and zebrafish whole genome shotgun data

In principal, mouse and zebrafish whole genome shotgun data was assembled over targeted regions of interest in the same way as ESTs were clustered. However, chromatograms as well as their interpreted sequences were available from trace repositories (http://trace.ensembl.org/). Chromatogram data includes quality information as well as the sequence. The Phrap assembly software utilises chromatogram quality data in the assembly process. To make use of this additional information, the automated retrieval of chromatograms was incorporated in to the iterative assembly process. The seed dataset was defined as for the EST clustering.

Subsequent iterative searching made use of the SSAHA (section 2.11.2) software with word length set to 28 for rapid identification of overlapping sequences.

Tetraodon nigroviridis whole genome shotgun data

Whole genome shotgun sequence data representing 2.5x genome coverage of the *Tetraodon nigroviridis* genome was obtained (H Roest Crollius, personal communication). The seed data set was defined by BLASTN searching (bitscore >60) with the repeat masked *Fugu* sequence contig (section 3.5.6). Iterative contig extension and assembly were then carried out as described for EST clustering. A novel, *Tetraodon* repeat that was not incorporated in RepBase and was consequently not recognised by RepeatMasker (sections 2.11.2 and 2.11.1), was manually masked during the assembly process.

Chapter 3

Cloning the *DISC1* region from *Fugu rubripes*

3.1 Preface

The compact nature of the *Fugu* genome (Hinegardner, 1968), a similar gene repertoire to other vertebrates (Brenner *et al.*, 1993) and a paucity of interspersed repetitive elements have led to the proposition of *Fugu rubripes* as a model vertebrate genome (Brenner *et al.*, 1993; section 1.7). Genomic sequence from a *Fugu DISC1* locus (if it exists and can be identified) would provide insight into the wider genomic organisation of the 1q42 breakpoint region and an opportunity for the identification of previously unidentified exons and non-coding regulatory sequences. It would also provide a platform for investigating the nature and conservation of the *DISC2* transcript and any of the syntenically conserved genes identified in *Fugu*.

The pattern of amino acid conservation in a *Fugu* homologue of *DISC1* would give important new insight into functional constraint and possibly structural features of this novel protein which currently has no known homologues (section 1.6.6). Investigations were therefore undertaken to identify a *Fugu* homologue of *DISC1*.

3.2 Hunting for *Fugu DISC1*

A combination of strategies were undertaken to predict structural or functional features of the DISC1 protein with the aim of identifying regions of likely evolutionary conservation to direct the selection of probes for use in cross species hybridisation experiments. Such probes would then be used to screen the *Fugu* genomic library detailed in section 2.9.7 and contribute to the cloning of *DISC1* and physically linked genes from other species.

3.2.1 Prediction of *DISC1* conserved regions

Public sequence databases were searched using BLASTP and TBLASTN (section 2.11.2) to identify homologues of human *DISC1*. The distribution of sequence matches along the length of DISC1 strongly suggested that the protein consists of two principal domains. The C-terminal 503 amino acids displayed weak similarity (up to 23% identity and 42% similarity) to myosins, tropomyosins and other proteins with extended regions of coiled coil. The N-terminal domain showed only very weak similarity to a wide variety of proteins, which was no longer reported when compositionally biased sequences were masked (BLASTP with SEG filtering, section 2.11.2), suggesting the similarity was principally due to compositional bias.

The hypothesis that the DISC1 protein is comprised of two principal regions was further supported by striking compositional differences between the N-terminal (amino acids 1 to 350) and C-terminal (351 to 854) regions. The N-terminal region is highly enriched for small amino acids (A, S and G) where as the C-terminal region is enriched for large amino acids (A, K, R, I, L and E), a distinction that can be clearly observed in figure 3.1 (track 'b').

Coiled coil prediction using the Coils program (section 2.11.2) reported strongly predicted coiled coil stretches in the C-terminal half of DISC1, consistent with the observed similarity to myosins and structurally related proteins. The output of Coils and Multicoil under multiple window sizes (section 2.11.2) is summarised in figure 3.1. TMPred (section 2.11.2) with default settings was used to predict transmembrane regions of DISC1: a single possible transmembrane region was reported corresponding to amino acids 50 to 68 of the human DISC1 reference sequence. However, the reported score of 909 does not strongly predict a transmembrane helix in the absence of supporting evidence such as the prediction of other such helicies within the same protein. No transmembrane helices were predicted by the TMap program (section 2.11.2). Searching against the PFAM (section 2.11.1) database of known protein domains failed to find any significant matches to DISC1.

Multiple secondary structure predictions were made, coordinated through the J-pred interface (section 2.11.2). Although secondary structure prediction is notoriously problematic, especially in the absence of multiple homologous sequences, predictions for DISC1 were consistent between methods. A consensus prediction of a helical only C-terminal region and an N-terminal region with interspersed helix and sheet structures was generated (figure 3.1). J-pred derived predictions of residue solvent accessibility also showed a striking distinction between the C-terminal and N-terminal regions, with no runs of more than 4 buried residues in the C-terminus and many stretches of 5 or more buried residues in the N-terminus (figure 3.1). This suggests that the N-terminal region could fold into a compact globular structure whereas the C-terminal region is expected to be a more open, solvent exposed structure. In agreement with these findings, Globe (section 2.11.2) prediction of protein globularity reports that the distribution of hydrophobicity within the N-terminal domain is consistent with it forming a globular structure. Within the C-terminal region, the distribution of hydrophobicity is not consistent with a compact globular structure.

The only strongly predicted, clearly defined motifs identified within the amino acid sequence for DISC1 were two leucine zipper motifs within the C-terminal region (figure 3.1). Leucine zippers consist of a heptad repeat of leucine residues for at least eight turns of an α-helix (Prosite:PDOC00029). Leucine zippers mediate protein-protein interactions through the formation of a coiled coil structure (see section 6.7.2 for a more detailed description) with a second leucine zipper. Leucine zippers are functionally important and highly conserved in Bzip-like transcription factors (Hurst, 1995), but outside these protein families there is no reason to believe that they are functionally distinct or more likely to be conserved than other regions of coiled coil. As DISC1 has no detectable homology to DNA binding domains, it was considered unlikely that the leucine zippers would show higher levels of cross species conservation than any other coiled coil region of DISC1.

The periodicity of amino acid physico-chemical properties is essential for coiled coil formation and for mediating the specificity of interactions (Beck and Brodsky, 1998;

Betz *et al.,* 1995). For solvent exposed coiled coils there is little steric constraint on amino acids, other than those at the dimerising interface (positions 1 and 4 in the heptad repeat). This lack of constraint generally leads to poor amino acid conservation of solvent exposed coiled coils. In contrast, the many buried residues in globular protein folds tend to have both physico-chemical and steric constraints. Based on the "globular head and coiled coil tail" model of DISC1 (figure 3.1), it was predicted that the N-terminal head region would therefore show a higher level of conservation than the C-terminal region of the protein.

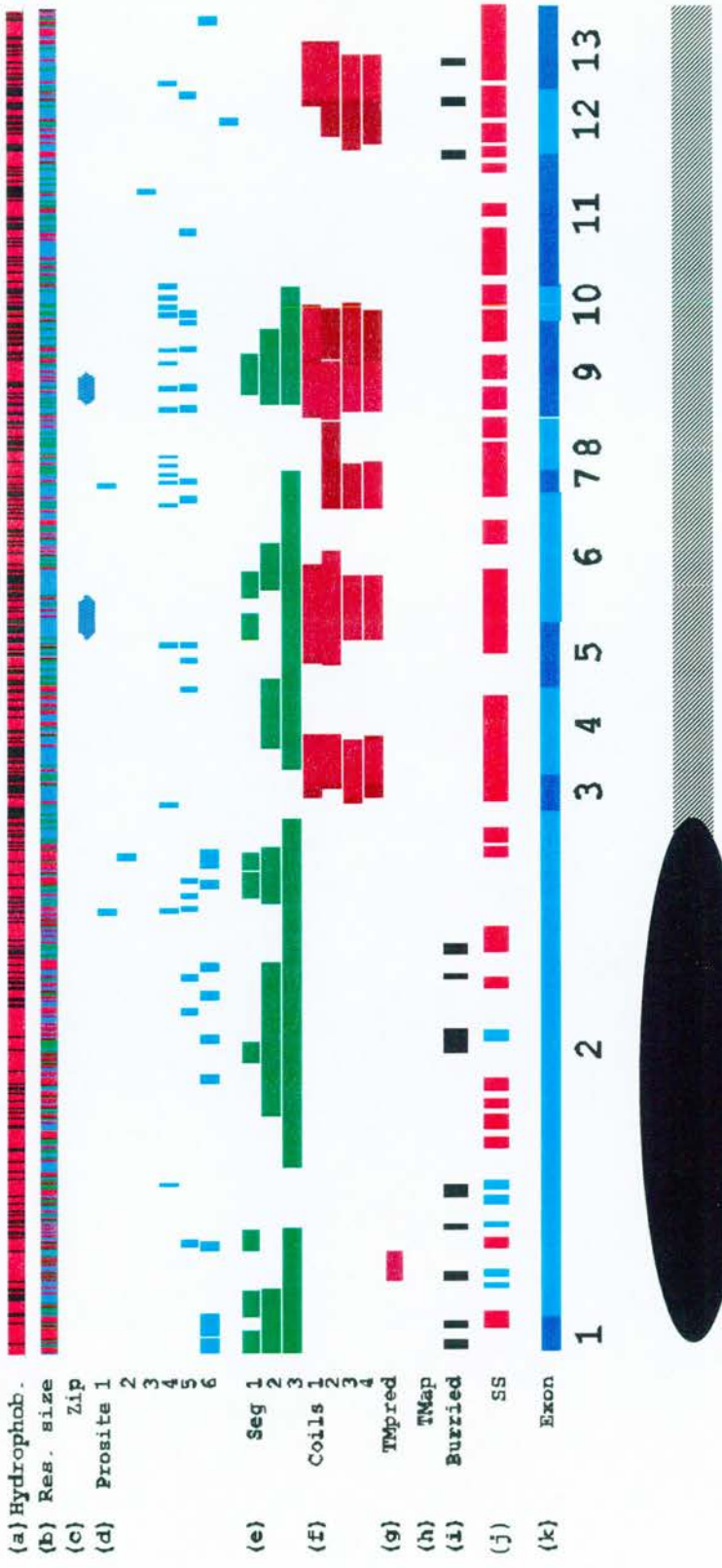


Figure 3.1; Summary of DISC1 features and predictions. See next page for full legend.

Figure 3.1; Summary of DISC1 features and predictions. **(a)** Indicates the relative hydrophobicity of sequence where red indicates hydrophobic residues. **(b)** "Res. size" shows the distribution of small (AGS, red), medium (NDCPTV, green) and large (REQHILKMFWY, blue) residues along the length of DISC1. **(c)** "Zip" Shows matches to the Prosite motif for leucine zippers, a specialised form of coiled coil. **(d)** Matches to Prosite patterns for ASN glycosylation, glycosaminoglycan, CAMP phosphorylation, protein kinase C phosphorylation, CK2 phosphorylation, and N-myristyulation modification sites respectively for 1 to 6. **(e)** Seg shows regions of DISC1 with reduced compositional complexity, Seg 1 is with default settings, Seg 2 and 3 use modified parameters of W=25, K1=3.0 and K2=3.3 for Seg 2 and W=45, K1=3.4 and K2=3.75 for Seg 3. Seg 1, 2 and 3 are progressively less stringent at defining low complexity regions of protein sequence. **(f)** Output from Coils (section 2.11.2) for the prediction of coiled coil regions. Two different matrices (MTK and MTIDK) are used both with and without weighting (section 2.11.2), Coils 1 is MTK, 2 is MTK weighted, 3 is MTIDK and 4 is MTIDK weighted. **(g)** Potential transmembrane regions as predicted by TMpred (section 2.11.2). **(h)** Potential transmembrane regions as predicted by TMap (section 2.11.2). **(i)** "Buried" indicates regions of DISC1 with 4 or more consecutive residues that are predicted (JPred2 consensus) to be buried (not solvent exposed). **(j)** "SS" indicates the consensus secondary structure prediction from JPred2 (section 2.11.2), with brown boxes indicating a consensus helix prediction and blue boxes indicating consensus beta sheet prediction. **(k)** The region of DISC1 amino acid sequence encoded by each of the 13 exons. **(l)** A generalised model for DISC1 structure. An N-terminal region composed of one or more globular domains with both α-helix and β-sheet secondary structure elements. The C-terminal region is composed entirely of α-helix and loop secondary structure with multiple regions of strongly predicted coiled coil. Features 'a' to 'h' were predicted through the PIX interface (http://menu.hgmp.mrc.ac.uk/PIX/), 'i' and 'j' through the JPred2 interface (http://jura.ebi.ac.uk:8888/submit.html).

3.2.2 Probe design

Probes for cross species hybridisation were initially designed from the 980 bp exon of human *DISC1* (subsequently referred to as exon 2). This exon encodes the majority of the predicted globular N-terminal region of *DISC1*. The whole of human *DISC1* exon 2 was used as a probe (subsequently referred to as D1P) to screen the available *Fugu* genomic library.

3.2.3 Genomic library screening

The MRC HGMP-RC distributed, total *Fugu* genomic cosmid library in gridded clone format (section 2.9.7) was screened by hybridisation (section 2.10.2). The first round of library screens used the D1P probe under low stringency conditions (section 2.10.2). Six double spotted clones (each clone was spotted twice onto a filter) gave detectable hybridisation signals: 184-F17, 184-G17, 184-G16, 183-G6, 22-E22 and 21-E22. The obvious clustering of clone names suggested that there had been cross-contamination of clones either in the original plates or in the spotting of the gridded filters. These six cosmid clones were obtained from the HGMP-RC, analysed by restriction analysis and re-hybridised with the original D1P probe (figure 3.2). Restriction analysis demonstrated that the six clones represented only three unique inserts, confirming that there had been cross-contamination of clones, probably in the original plates.

The approximately 1.6 kb *BamHI* fragment of clone 184-G16 that gave a specific hybridisation signal to the D1P probe was cloned into the pBluescript II SK- vector (sections 2.2.4). Hybridisation of D1P to the sub-clone confirmed the appropriate band had been isolated (figure 3.2). Sequencing the ends of the clone using vector primers, identified sequence similarity to a known human gene (KIAA0389) and to *Fugu* genomic sequence that was being generated from clone 184-G17 by Greg Elgar (HGMP-RC, UK). Further sequences from the cosmid clone 184-G17 (identical by restriction analysis to 184-G16) generated at the HGMP-RC, showed substantial similarity to hsp70-2 and an unnamed immunoglobulin gene. At that time, both human hsp70-2 and the immunoglobulin gene were mapped to the MHC locus on human chromosome 6. The gene KIAA0389 has also, subsequently been mapped to

this locus (ENSEMBL 1.0). Sequence of this 1.6 kb *Fugu* genomic fragment did not share significant similarity with human DISC1 amino acid sequence (not shown).

In the experiments described above, hybridisation and washing were carried out at low stringency, as the degree of homology between human and *Fugu DISC1* sequences was unknown. Subsequently a modified strategy was adopted, which avoided reliance on a single probe under the necessarily low stringency hybridisation conditions. Multiple hybridisations were carried out and only clones that gave strong hybridisation signals with a single probe or weak signals with multiple probes were investigated further. In addition, failure to identify *DISC1* homologous sequences in the *Fugu* genomic library suggested that previous prediction of N-terminal conservation (section 3.2) may not be valid. A second generation of probes was therefore designed to both the N-terminal and C-terminal regions of human *DISC1* (appendix I).

The low stringency hybridisation and washing conditions (section 2.10.2) resulted in large numbers of weakly positive clones. Autoradiographs of successive hybridisations with non-overlapping probes were aligned and those clones showing hybridisation signals for more than one independent probe were investigated further. It is important to note that between hybridisations, filters were stripped of bound probe (section 2.10.9) and checked by autoradiography for residual signal prior to re-probing.

In these second generation screens, no clones were considered strongly positive for any probe, but 43 cosmid clones were identified as weakly positive for multiple probes. Each of these clones were further characterised by restriction digestion, Southern blotting and hybridisation with the probes that originally gave weakly positive signals. Of the 43 clones, only five produced convincing signals in this re-analysis. For each of the clones the hybridising restriction fragment was sub-cloned and sequenced. None of these sub-clones showed significant similarity to human *DISC1* nucleotide or amino acid sequence (data not shown).

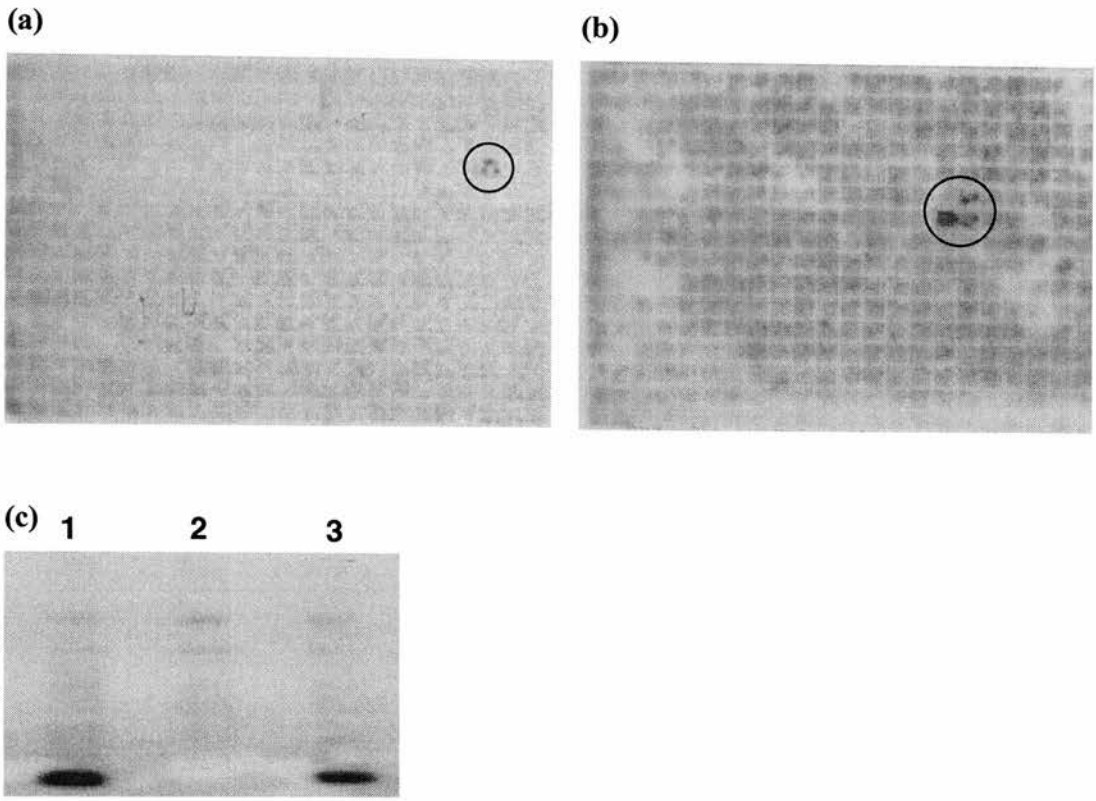


Figure 3.2; Screening *Fugu* genomic cosmid libraries for a homologue of *DISC1* using the human derived probe D1P. The library filters were gridded and double spotted as described in section 2.9.4. Only 2 clusters of hybridisation signal above back ground were found. **(a)** Library filter 1, panel 3 showing 2 hybridising clones (21-E22 and 22-E22) in close proximity. **(b)** A further 2 clear hybridisation signals were found in close proximity to one another on Library filter 4, panel 5 (circled). There is a further undefined signal over the adjacent block of the grid. This pattern of hybridisation was interpreted as 184-F17, 184-G16 and 184-G17 although 184-G16 does not hybridise with the expected discrete double-spot pattern. **(c)** Restriction digests (*BamHI*) of cosmid clone 184-G16 after blotting and hybridisation with probe D1P. Single bands were cut from the gel prior to blotting, the 1.6 kb band cut from lane 2 was subsequently sub-cloned and sequenced.

3.2.4 cDNA library screening

The *Fugu* cosmid library screened is estimated to be 7.5 times genome coverage (section 2.9.7), therefore on average every nucleotide of the genome is expected to be represented 7 to 8 times. However, nucleic acid libraries are often biased in their representation due to the method of construction, local features of the nucleotide sequence or epigenetic modifications to the nucleotides (Wong *et al.,* 1993). As a complementary strategy to genomic library screening, 11 *Fugu* cDNA libraries distributed by the UK HGMP-RC (section 2.9.7), were screened using human derived *DISC1* probes (D1P, D79P, D80P and D95P; appendix I) under the low stringency conditions described for screening the *Fugu* cosmid library. All 11 libraries were screened in parallel, to facilitate direct comparison of signal strength between libraries and minimise reagent expense.

There were several potential advantages to using cDNA libraries over genomic cosmid libraries. The cDNA libraries were constructed in the high copy number pBlusescript II KS- vector (section 2.2.4) with up to several hundred copies of plasmid per bacterial cell, whereas the genomic library was constructed in the Lawrist4 cosmid vector (section 2.2.4) with less than 8 copies per cell. Therefore, the proportion and total quantity of cloned DNA present on the filter will be more than ten fold greater for cDNA clones than genomic cosmid clones. Hybridisation signals should subsequently be at least 10 fold stronger for cDNA library filters compared to genomic filters under identical conditions. *DISC1* is anticipated to be a transcribed and translated gene and cDNA libraries by their very nature are specifically enriched for the transcribed proportion of a genome. However, genes that are expressed at low levels or in highly restricted temporal or spatial patterns are likely to be under represented in these libraries.

Twenty four *Fugu* cDNA clones showed weak hybridisation to the D1P probe. Each was obtained from the HGMP-RC, plasmid DNA isolated and partially sequenced. Of the 24 clones isolated, none had the potential to encode a recognisable homologue

of *DISC1*. There was no obvious consistency of sequence motifs present amongst the clones identified.

3.2.5 Discussion - A dead end?

The repeated failure to isolate *Fugu* genomic or cDNA clones containing *DISC1* implies one of three scenarios.

(1) There is not a homologue of *DISC1* in the *Fugu* genome, either because the gene was not present in the last common ancestor of humans and *Fugu* or it was present in the last common ancestor, but was subsequently lost in the lineage to *Fugu*.

(2) There is a homologue of *DISC1* in the *Fugu* genome, but that sequence is not represented in the cosmid or cDNA libraries screened.

(3) There is a *DISC1* homologue in the *Fugu* genome that is represented in the libraries screened, but there has been sufficient sequence divergence since the last common ancestor for cross species hybridisation to fail, even under the low stringency conditions being used.

Scenario 1 would require the selection of an alternative model genome for the comparative investigation of the *DISC1* region and would likely rule out use of Tetraodontiform species as a whole, with all of the associated advantages of this group of organisms for comparative genomic studies (section 1.7.1). Scenario 2 could be distinguished from 1 and 3 by *Fugu* total genomic restriction digestion, Southern blotting and hybridisation with the human *DISC1* probes. To address this issue, total *Fugu* genomic DNA was digested with *EcoRI*, Southern blotted and hybridised with the D1P probe. No signal was detected (data not shown), eliminating scenario 2. Without an alternative *Fugu* library to screen or further information on the *DISC1* genomic region in humans, this work could not continue.

3.3 Hunting for *Fugu TRAX*

3.3.1 *TRAX* – a new opportunity

Through the rapid amplification of cDNA ends (RACE) method, human *DISC1* was observed to intergenically splice with the TRanslin Associated factor X (*TRAX*) gene (Millar *et al.,* 2000). While it is conceivable that this splicing could be in *trans*, i.e. *TRAX* is not directly upstream of *DISC1* on chromosome 1, *trans* splicing has never been described in humans although it is common place in *C. elegans* (Blumenthal 1995 for review), and has been reported in *D. melanogaster* (Labrador *et al.,* 2001; Dorn *et al.,* 2001). Millar *et al.,* used high stringency nucleotide hybridisation of a *TRAX* cDNA to digests of human PAC clones from the chromosome 1 breakpoint contig (section 1.6.6), to confirm that *TRAX* is upstream of human *DISC1*. By inference from the intergenic splicing, *TRAX* was also assumed to be in the same transcriptional orientation as *DISC1*.

TRAX was first identified as a translin binding protein in yeast two hybrid assay (Aoki *et al.,* 1997) and has subsequently been characterised as a component of several molecular complexes that also contain translin. For details of *TRAX* function, biochemistry and evolution, see chapter 8. Through restriction and hybridisation analysis, K. Millar (Medical Genetics Section, University of Edinburgh) estimated the 3' end *TRAX* to be approximately 50 kb from the 5' end of *DISC1*. Unlike *DISC1*, *TRAX* has readily identifiable homologues in the genomes of *Drosophila melanogaster*, *Schizosacharomyces pombe*, *Arabidopsis thaliana* (figure 3.3) and is well represented in vertebrate EST collections (chapter 7).

One of the reasons for selecting *Fugu* as a model genome was the demonstration of conserved synteny between *Fugu* and other vertebrates (section 1.7.1). While gene order is not always conserved between vertebrates (Gilley *et al.,* 1997; Gilley *et al.,* 1999) it is highly likely that two neighbouring genes in human will also be adjacent in *Fugu*, or at least syntenic with localised gene order and orientation changes. McLysaght *et al.,* 2000 have estimated that 40-50% of gene pairs in *Fugu* maintain

synteny within the human genome. This is likely to be an underestimate (section 1.7.1 and McLysaght *et al.*, 2000). Therefore, the identification of a *Fugu* homologue of *TRAX* may provide the means to identify a *DISC1* homologue if synteny is conserved.

3.3.2 *TRAX* probe design

Strongly conserved regions of the *TRAX* gene were readily identified through the depth of sequence and range of evolutionary distances represented by the available *TRAX* orthologues (Chapter 8 and figure 3.3). Non-overlapping probe sets T-P1 / T-P3 and T-P2 / T-P4 (appendix I) were designed from the conserved regions indicated in figure 3.3 and used to identify *Fugu TRAX* through cross species hybridisation.

In addition to these probes, degenerate oligonucleotides (U644, U647 and U649) were designed to conserved sequence motifs within the coding sequence of *TRAX*, as summarised in figure 3.4. Inosine bases were incorporated at most of the sites discrepant between human and mouse sequence as inosine has the ability to form hydrogen bonds with all four of the standard bases in DNA.

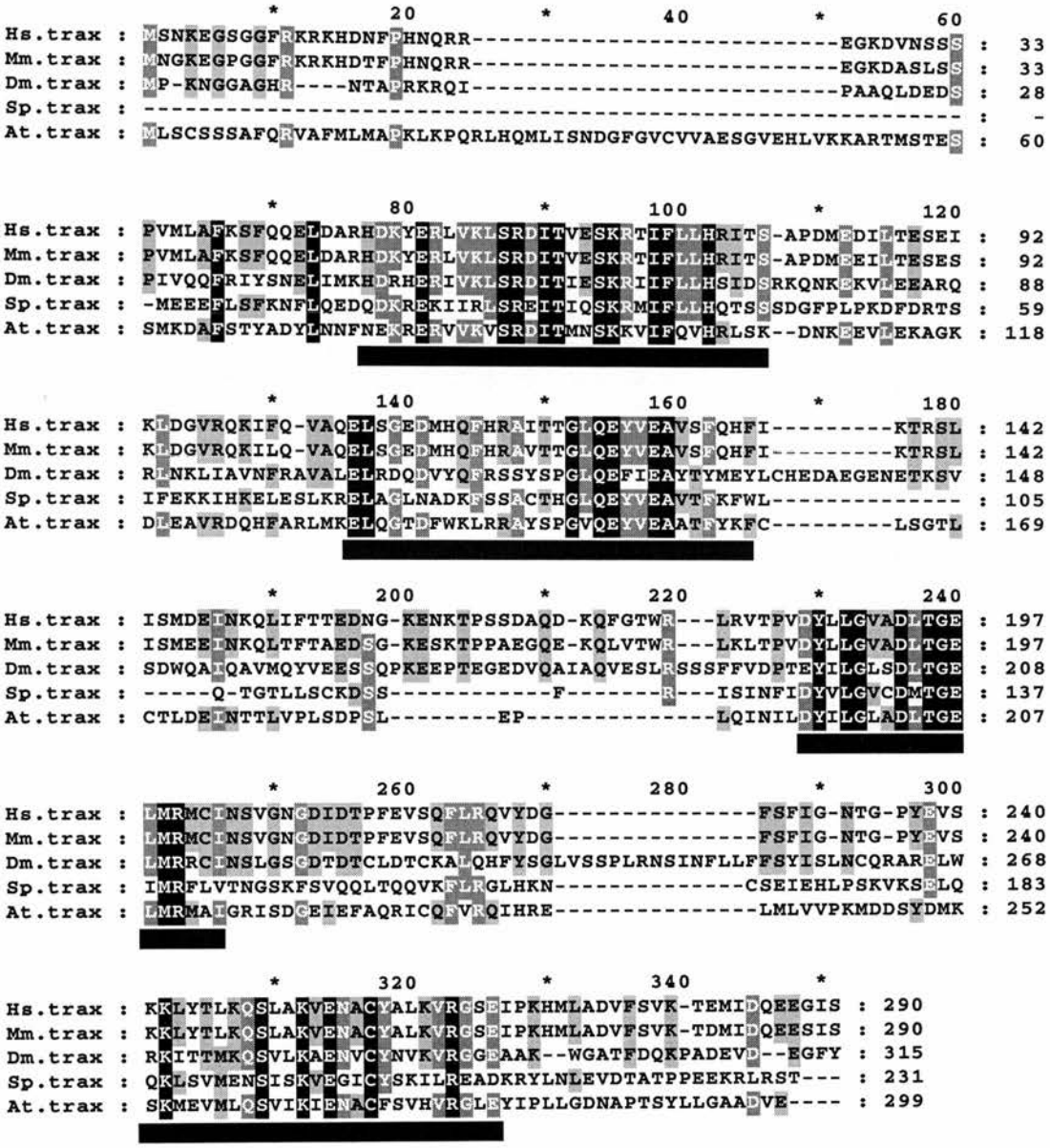


Figure 3.3; Multiple sequence alignment of *TRAX* protein orthologues. Abbreviatons: Hs, human (X95073); Mm, mouse (Q9QZE7); Dm, *Drosophila melanogaster* (Q9VF77); Sp, *Schizosacharomyces pombe* (AL02375); At, *Arabidopsis thaliana* (AL006836). Positions at which all residues are conserved have a black background and white foreground, positions with 4/5 conserved have dark grey background and 3/5 conserved have a light grey background and black foreground. Only amino acid identity is shown in this alignment (similarity groups are not represented). Horizontal black bars under the alignment define the highly conserved regions of *TRAX*.

```
Mm 564 ACAGGACTACAGGAATATGTGGAAGCTG 591
Hs 524 ACAGGACTGCAGGAATACGTGGAGGCTG 551
U644       GCACTICAGGAATAIGTGGAIGC


Mm 359 TGAGAGACTCGTGAAGCTAAGCCGGGATATTACTGTGGAGA 399
Hs 319 TGAGAGACTTGTGAAACTTAGTCGGGATATAACTGTTGAAA 359
U647     GAGAGACTIGTGAAICTIAGICGGGATATAAC


Mm 378 AGCTAAGCCGGGATATTACTGTGGAGAGTAAGAGGACAATTTTTCTTCTCCATA 424
Hs 338 AACTTAGTCGGGATATAACTGTTGAAAGTAAAAGGACAATTTTTCTCCTCCATA 384
U649         AGTCGGGATATAACTGTTGAAAGTAAAAGGACAATTTTTCTCCTCCA
```

Figure 3.4; Degenerate oligonucleotide probes designed to the most conserved region of the human *TRAX* open reading frame. "I" indicates incorporation of the base inosine. Mm indicates mouse *TRAX* cDNA sequence (EMBL: AF187040) and Hs indicates human *TRAX* cDNA (X95073). Numbers associated with sequences refer to the sequence coordinates shown in the alignment. The full length and sequence of each oligonucleotide is shown, oligonucleotide names begin U64.

3.3.3 Genomic library screening

The human *TRAX* probes summarised in appendix I were used to screen the *Fugu* genomic cosmid library in the manner described previously for human DISC1 probes (section 3.1.3). Hybridisation of T-P2 under low stringency conditons identified 19 weakly positive cosmid clones (Filter 1: 21-N12, 21-N14, 45-C01, 33-K19; Filter 2: 67-D19, 74-D08, 72-O22, 72-P21, 72-P22, 93-M01, 94-M01; Filter 3: 106-D05, 105-L05, 104-J18, 112-E17, 112-M06, 127-K19, 137-D19, 139-K17; Filter 4: 154-D3). Under the same hybridisation and washing conditions probe T-P1 did not give convincing hybridisation results. T-P4 hybridisation identified the same clones as T-P2 and other clones were weak positives. However, as T-P2 is completely contained within the sequence of T-P4 these results cannot be considered independent verification of positive signals.

Hybridisation of T-P3 to genomic filters under the same conditions as T-P1, 2 and 4 resulted in a signal pattern typical of an interspersed repetitive element (figure 3.5).

Re-sequencing of the T-P3 probe confirmed it represented the expected sequence, which did not contain a repeat. Searching the sequence of T-P3 against all known repetitive sequence and all publicly available *Fugu* sequence did not find any significant homology to repetitive elements that could explain this pattern of hybridisation. Hybridisation at 60°C rather than 50°C used in low stringency hybridisations resulted in the loss of all signals (data not shown).

An equi-molar pool of degenerate oligos U644, U647 and U649 was end labelled (figure 3.4) and hybridised to the *Fugu* genomic (cosmid) filters under low stringency conditions. Of 116 cosmid clones giving identifiable hybridisation signals to the pool of oligonucleotides, five had previously been shown to hybridise to probe T-P2. As there is no sequence overlap between any of the 3 oligonucleotides and T-P2, these clones were considered to be "double positive" and were subsequently followed up.

3.3.4 Clone evaluation

The five double positive clones (21-N12, 21-N14, 33-F19, 72-O22 and 108-L5) were obtained from the UK HGMP-RC. Clones were streaked onto appropriately selective agar plates and single colonies picked to ensure single clone populations. Cosmid DNA was prepared from each clone (section 2.2) and digested with the restriction endonuclease *EcoRI* (figure 3.6). Based on restriction analysis, clones 21-N12, 21-N14, 108-L5 and 72-O22 showed a substantial level of overlap. 21-N12 and 21-N14 had identical restriction patterns and given the close proximity of their coordinates it was likely that they represented the same clone (one clone contaminating an adjacent well). However, clone 108-K19 shares only the 1.24 kb and 3.76 kb cosmid vector bands with the other clones, suggesting that it was unrelated to the other 4 clones.

Southern blotting (section 2.9.1) of the restriction digest followed by hybridisation of the blot with the human *TRAX* probe, T-P4 resulted in a strong and specific hybridisation signal in the clones 21-N12, 21-N14, 33-F19 and 72-O22 at the position of the 5 kb *EcoRI* fragment (figure 3.6).

The 5 kb *EcoRI* restriction fragment was cloned into the pBluescript SK- cloning vector (section 2.2.4). Transformants were screened by hybridising with the T-P4 probe (figure 3.6). The inserts of T-P4 positive sub-clones #38 and #45 were sequenced in their entirety, initially with vector primers and subsequently with "walking" oligonucleotides designed to the ends of sequence as it was produced. Preliminary analysis (BLASTX against the SPTR database) of the sequence within sub-clones #38 and #45 demonstrated that a *Fugu* homologue of human *TRAX* had been identified (figure 3.7). The isolation and sequencing of a *Fugu TRAX* cDNA clone (detailed in section 6.3.2) also allowed oligonucleotides to be designed from predicted exon sequences (section 6.3.2) to initiate the walking process at internal locations, expediting completion of the sequencing effort. Complete sequencing of the 5 kb insert was achieved in eight walking steps.

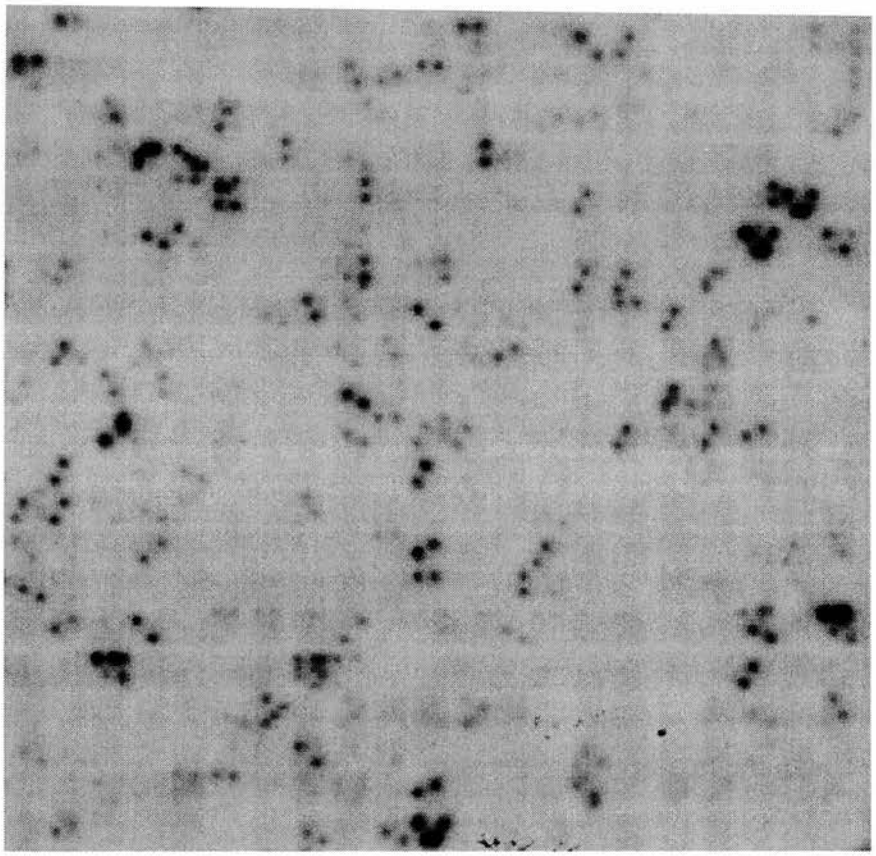


Figure 3.5; Hybridisation of human *TRAX* probe T-P3 to one of the four *Fugu* genomic cosmid library gridded filters. The high number of positives suggested that T-P3 hybridises to an interspersed repetitive element.

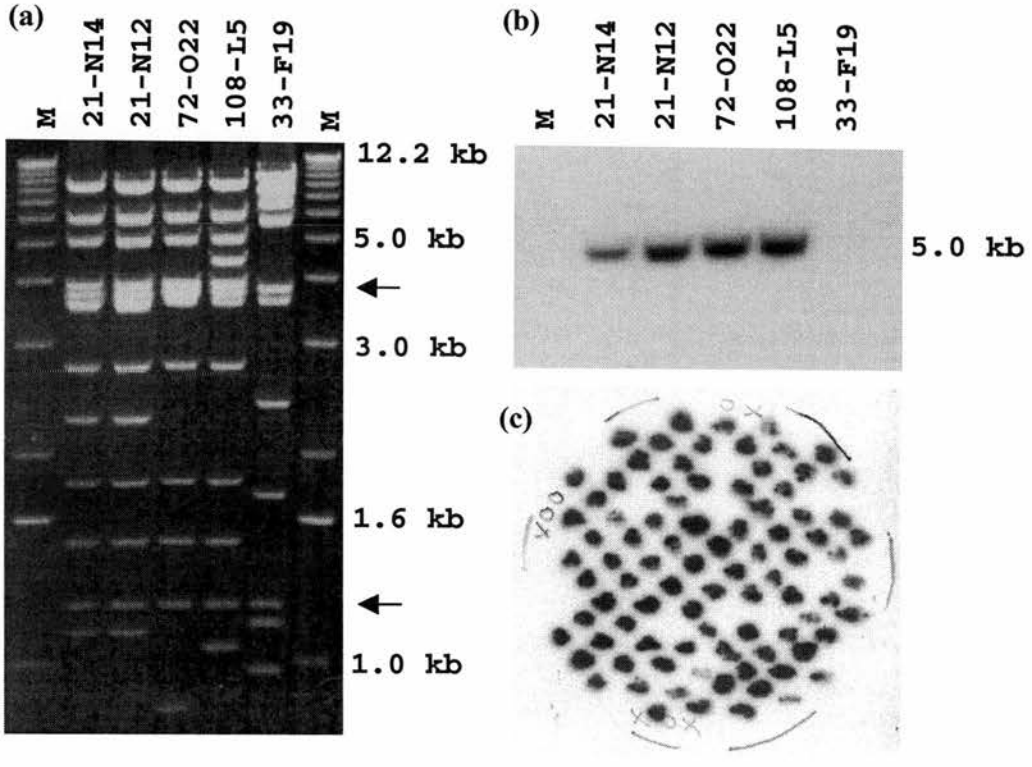


Figure 3.6; Evaluation of *Fugu* cosmid clones. **(a)** *EcoRI* digests of the candidate *TRAX* containing clones. M denotes marker lanes using the 1 kb marker (section 2.7.2). The gel was 0.8% agarose with 1× TBE buffer (section 2.7). Arrows indicate restriction fragments derived from the cosmid vector Lawrist4. Clones 21-N14, 21-N12, 72-O22 and 108-L5 had similar restriction patterns, clone 33-F19 appeared to be unrelated. **(b)** Autoradiograph of T-P4 probe hybridised to a Southern blot (section 2.9.1) of the gel shown in section 'a'. A clean and strong hybridisation signal was observed, corresponding to a 5 kb *EcoRI* restriction fragment present in all clones except 33-F19. **(c)** Hybridisation of T-P4 to 5 kb EcoRI fragment transformants identifying multiple positive clones. Marks (xoo, ox and xox) on the autoradiograph were used to align the autoradiograph back to a replicate plate for the selection of positive clones.

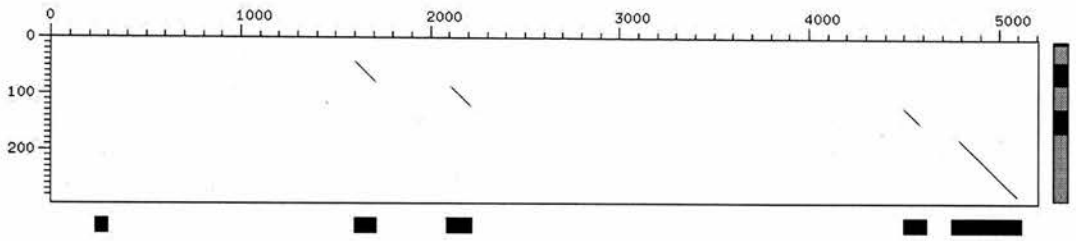


Figure 3.7; Alignment of human *TRAX* amino acid sequence against that of the *Fugu* 5 kb *EcoRI* fragment. Human *TRAX* amino acid sequence is represented on the vertical, the sequence of the 5 kb fragment along the horizontal. The alignment was produced using dotter (section 2.11.2) with a window size of 17. The vertical bar to the right of the alignment shows alternate black and grey boxes for coding sequence derived from each exon of human *TRAX*. Black boxes below the alignment show genscan prediction of exons (section 2.11.2) within the approximately 5 kb fragment of sequence.

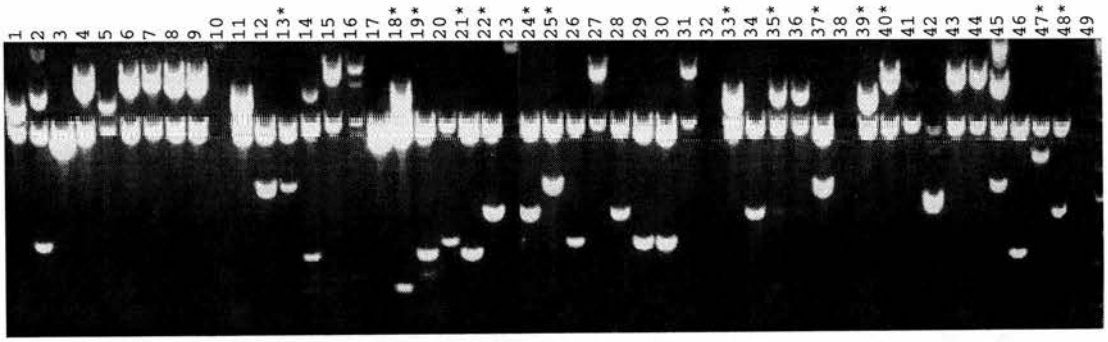


Figure 3.8; Sub-clones of the *Fugu* cosmid 21-N14 digested by *EcoRI* to reveal the size of insert fragments. Preparations 10, 23, 32, 38 and 49 failed. The restriction pattern of clones was used to determine a minimally redundant set of clones for sequence scanning. Those clones marked with a '*' were selected for insert end sequencing.

3.4 Contig assembly and extension

3.4.1 Sequence sampling

To identify other genes within the *Fugu TRAX* containing contig, clone 21-N14 was digested to completion with *EcoRI* and a sub-clone library generated in the pBluescript II SK- vector (section 2.2.4). Sizes of the sub-clones were obtained by *EcoRI* restriction digestion identifying a minimally redundant subset for sequencing (figure 3.8). The ends of each selected sub-clone were sequenced using vector primers. Cosmid insert ends were also sequenced directly on cosmid preparations.

3.4.2 *DISC1* and other genes

Sequences from cosmid 21-N14 showed similarity to the human genes nidogen (*NID*), Transmembrane-7-subfamily-1 (*TM7SF1*) and crucially *DISC1* (figure 3.9). The approximate position of these genes could be determined (figure 3.9) using the restriction map, insert sizes of sub-clones from which end sequence had been generated and the sequence generated directly from the cosmid ends. The genes *NID* and *TM7SF1* are discussed in more detail in section 4.2.2.

3.4.3 Extending the contig

Homology to *DISC1* exons 8 and 9 was identified in sequence that could be located and oriented at the right hand side of the current cosmid contig (figure 3.10), approximately 1 kb from the end of the contig. Human *DISC1* is known to have at least 13 exons (Millar *et al.,* 2001a) and there is at least one PAC length (more than 100 kb) between exons 9 and 13 (K. Millar, personal communication). At least a 100 fold genomic compaction of *Fugu DISC1* would therefore be required for the orthologue of exon 13 to be contained in the current contig, and since *Fugu* genes are typically 7.5 to 8 fold compact relative to their human orthologue it was considered unlikely that the current cosmid contig would contain the complete *DISC1* gene. The 2.1 kb insert of sub-clone #47 that contained homology to human *DISC1* exons 8 and 9 was used as a hybridisation probe at high stringency to screen the *Fugu* genomic cosmid library to identify clones to extend the existing contig towards the 3' end of *DISC1*. As expected, all 4 of the previously identified cosmid clones gave strong

hybridisation signals to the insert of sub-clone #47. Additionally clones 33-I19 and 5-F14 gave equally strong hybridisation signals.

Clone 33-I19 was located adjacent to 33-F19 on the gridded filters. 33-F19 was previously investigated after giving a putative hybridisation signal to human *TRAX* sequences (section 3.3.3), but was found to be unrelated to other identified clones and failed to re-hybridise the original probe. Re-examination of the original filters demonstrated that it was clone 33-I19, not 33-F19 that gave the original hybridisation signals. The absence of background signal in the original *TRAX* hybridisation caused autoradiographs to be misaligned to the gridding pattern and the coordinate to be misidentified.

Cosmid clone 5-F14 had not been identified by *TRAX* probes, suggesting that it would extend the existing cosmid contig towards the 3' end of *DISC1*. Cones 33-I19 and 5-F14 were obtained from the HGMP-RC and analysed by restriction analysis demonstrating that both 33-I19 and 5-F14 extended the existing cosmid contig (figure 3.10).

To further extend the cosmid contig, probe F14E (coordinates 44047 - 44504 of the sequence reported in section 3.5.6) located in sequence downstream of the *DISC1* gene in cosmid 5-F14, was hybridised at high stringency to the gridded cosmid library. F14E failed to hybridise to any clones in the cosmid library other than 5-F14. A *Fugu* genomic BAC library (IncyteGenomics BAC-6253) was also screened (in collaboration with Greg Elgar and Phil Snell, UK HGMP-RC) with the F14E probe and a probe derived from *Fugu DISC1* cDNA sequence (section 6.3.2). No clones in the BAC library gave a positive hybridisation signal. This lack of sequence coverage in both libraries suggests that sequence downstream of *DISC1* may be unstable in *E. coli* or that there is a cloning bias in the libraries screened. A cloning bias has been reported for the *Fugu* whole genome shotgun sequencing effort, favouring the cloning of the 5' ends of genes (G. Elgar, personal communication).

(a) Similarity to human nidogen

```
HSP 1

Query: 337 QKNTFQCVFASMETASFAVLLNVNNGLQF--TSIDAGTSAAVMHAGFSKGLVVGFLFSSQ 164
           ++NTFQ V AS +++S+A+ L   +GLQF  T      +       FS+G  VGFL+ S
Sbjct: 175 KRNTFQAVLASSDSSSYAIFLYPEDGLQFHTTFSKKENNQVPAVVAFSQG-SVGFLWKSN 233

Query: 163 GPYYRITTDEEDSIRALAE*DHWGLGG 83
           G  Y I  ++ +SI  LA+  + G  G
Sbjct: 234 GA-YNIFANDRESIENLAKSSNSGQQG 259


HSP 2

Query: 47  LSRETNSGMQGVWVY 3
           L++ +NSG QGVWV+
Sbjct: 249 LAKSSNSGQQGVWVF 263
```

(b) Similarity to human *TM7SF1*

```
HSP 1

Query: 747 QGLF*AKSKVAPQLLRFR*AGGCLGACVAPARVCFPCGVFAAAGCPMYLLFLFVSLLFLV 568
           Q +F AKSK +P+LL++R                              P+YL  LF+SL+FL+
Sbjct: 123 QVIFKAKSKYSPELLKYR--------------------------LPLYLASLFISLVFLL 156

Query: 567 VNLVCALLVRTSSTDTHAIVLVRVAINDTLFVLSAVSLSVCLYKIAKMSLANIYLESKVR 388
           VNL CA+LV+T   D   IV VRVAINDTLFVL A+SLS+CLYKI+KMSLANIYLESK
Sbjct: 157 VNLTCAVLVKTGDWDRKVIVSVRVAINDTLFVLCAISLSICLYKISKMSLANIYLESKGS 216

Query: 387 SV--YPNVGLSLILLCSS 340
           SV     +G+++ILL +S
Sbjct: 217 SVCQVTAIGVTVILLYAS 234


HSP 2

Query: 296 QGTSVCQVTGVGAVVILLYSSRACYNLVVLGLSN-KSINSFDYDWYNVSDQVRRRS 132
           +G+SVCQVT +G  VILLY+SRACYNL +L  S  K+++SFDYDWYNVSDQ   +S
Sbjct: 214 KGSSVCQVTAIGVTVILLYASRACYNLFILSFSQIKNVHSFDYDWYNVSDQADLKS 269
```

(c) Similarity to human *DISC1*

```
Query: 160 SLPPYPPQVVMSQRLGSSLRRKVSETETQLLALQEAKLAAISGNPALPVT*SRRY*MRIG 339
           SL     +V MS++  S+LR+KV++ ETQL AL EAK+ AIS
Sbjct: 556 SLKEITTKVCMSEKFCSTLRKKVNDIETQLPALLEAKMHAIS------------------ 597

Query: 340 ENLWALGEPGAGNDFSSAKELKAEMKGVHQERERLEALATRLQSLSS 480
                    GN F +AK+L  E++ +  ERE LE L ++L  LSS
Sbjct: 598 ----------GNHFWTAKDLTEEIRSLTSEREGLEGLLSKLLVLSS 633
```

Figure 3.9; Nidogen, *TM7SF1* and *DISC1* homology. BLASTX (section 2.11.2) searches carried out for all fragments of sequence generated through sequence sampling (section 3.4.1). Searches were against the SPTR database (section 2.11.1) with SEG filtering swiched off. **(a)** Similarity to human nidogen in two BLASTX high scoring segment pairs (HSP). **(b)** Similarity to human *TM7SF1* in two HSPs but likely representing three exons in *Fugu*. **(c)** Similarity to human *DISC1* in a single HSP. This amino acid sequence corresponds to exons 8 and 9 of human *DISC1*.

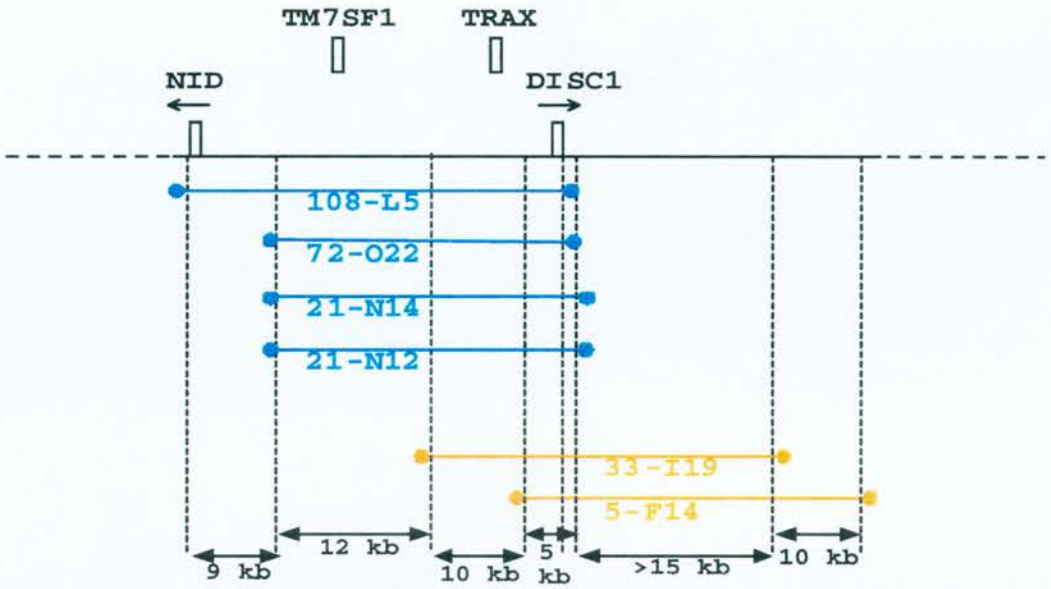


Figure 3.10; Map of the *Fugu* cosmid contig. The contig was constructed by sequence sampling (section 3.4.1), cosmid clone restriction analysis (section 3.3.4) and Southern hybridisation of restriction digests (section 3.3.4). Cosmids indicated as blue dumbbells were isolated in the first screen (section 3.3.3). Cosmids indicated as orange dumbbells were isolated in the second screen using *Fugu DISC1* derived sequence (section 3.4.3). Homology to four known human genes was identified from sequence scanning and direct sequencing of cosmid ends. The order and transcriptional orientation of *NID* and *DISC1* were unambiguously assigned (section 3.4.2) and are indicated by arrows. The position of *TM7SF1* and *TRAX* could only be approximately assigned on the cosmid contig (section 3.4.2), although the relative position of genes was unambiguously determined.

3.5 The *Fugu* contig: A sequencing project

The region of primary interest for comparative investigation was the *TRAX – DISC1* genomic region. Preliminary mapping evidence suggested that nidogen was at least several megabases downstream of *DISC1* in humans (Olsen *et al.*, 1989) and is therefore less likely to be relevant to the phenotype in the t(1;11) family. Nidogen is upstream of *DISC1* in *Fugu* suggesting a disruption of conservation in gene order and orientation between humans and *Fugu*. Cosmid clone 33-I19 was known to contain the whole of *TRAX* and was expected to contain at least exons 1 to 9 of *DISC1* (figure 3.9 and 3.10). Cosmid 5-F14 substantially overlaps with 33-I19 and extends the contig approximately 10 kb further downstream of *DISC1* than 33-I19. For these reasons cosmid 33-I19 was selected for compete sequencing and 5-F14 for targeted sequencing.

3.5.1 Sequencing strategy

The 30 to 40 kb inserts of cosmids are too large to be efficiently sequenced in a sequence walking approach as described for the *TRAX* containing 5 kb *EcoRI* fragment (section 3.3.4). Through the generation of sub-clone libraries, large clones are broken into many fragments which can then be sequenced in a high throughput manner using vector derived, standard oligonucleotides to prime sequencing reactions (section 2.2.4).

Fragmentation of large clones is typically achieved through sonication or restriction digestion, followed by ligation of the fragments (section 2.5.1) into an appropriate plasmid vector (section 2.2.4). Sonication shears DNA through mechanical stress, causing double strand DNA breaks randomly along the length of the molecule. The average size of fragments generated can be regulated by the length and intensity of sonnication. Restriction digestion (section 2.5.5) results in the specific enzymatic cleavage of DNA at (or near for some enzymes) the recognition site of a specific restriction endonuclease. As the cleavage is sequence dependent the distribution of sites along a DNA molecule is often biased. This bias is generally undesirable for the construction of sub-clone libraries and can be overcome in part by using a range of restriction enzymes.

Sonication is the method of choice in large scale sequencing projects (Venter *et al.*, 2001; International Human Genome Sequencing Consortium 2001; Adams *et al.*, 2000) however, during the ligation of inserts into vectors, inserts can join together prior to ligation into the vector and form a "chimeric" clone. Such clones are not readily detected until there is substantial sequence coverage across the region, at which point further experimentation is required to distinguish the genuine sequence from the chimeric sequence. Methods such as a dephosphorylation step (section 2.5.6) prior to ligation can be used to reduce the frequency of chimerism, but do not eliminate it.

For the sequencing of cosmid clones 33-I19 and 5-F14, a restriction based strategy for sub-clone library generation was deemed the most appropriate. Although sub-cloning by this method is subject to bias, the relatively small scale of the project allowed restriction enzymes to be pre-selected on the basis of how frequently they cleave the cosmid clones and the distribution of fragment sizes (section 3.5.2). The problem of chimerism still exists using this strategy, but as complete restriction digestion is being used (every restriction site in every target molecule should be cleaved), any insert fragments that contain the restriction site used in library construction can be considered to be chimeric and their sequence treated with appropriate caution during assembly.

3.5.2 Sub-clone library generation

High quality preparations (section 2.2.7) of cosmid 33-I19 were digested with the restriction enzymes *EcoRI, BamHI, NsiI, PstI, XhoI* and *XbaI* (all have 6 bp recognition sites and produce sticky ended products (section 2.5.5)). Products of these digests were separated on agarose gels and evaluated by average fragment size and range of fragment size (not shown). On this basis *EcoRI* and *PstI* were selected for production of sub-clone libraries. In addition to these "large" sub-clones of 1.5 to 9 kb, the restriction enzyme *Sau3AI* (4 bp recognition site) was selected to produce a library of small fragments in the 0.5 to 1.5 kb size range.

For each of the enzymes *EcoRI*, *PstI* and *Sau3AI*, 3.4 µg of cosmid 33-I19 was digested to completion. Digestion products were size selected by cutting from LMP agarose gel (section 2.4.4). Size selection was used in the construction of all libraries as small fragments insert into vectors preferentially to larger fragments. DNA was recovered from the gel slices and ligated into appropriately prepared pBluescript II SK- vector. Single transformants were selected and stored in 96 well plates. One plate of *EcoRI*, one plate of *PstI* and two plates of *Sau3AI* transformants were made. Similarly, an *EcoRI* library of cosmid 5-F14 was constructed. From the restriction digests of 5-F14 and 33-I19 it was apparent that the majority of sequence unique to 5-F14 was contained in a single approximately 17 kb *EcoRI* fragment. This *EcoRI* fragment was isolated (section 2.4.4) and digested with *Sau3AI* to produce a targeted 5-F14 library.

3.5.3 Sub-clone library validation

Restriction digestion of every clone in the sub-clone libraries was carried out to assess the complexity of libraries generated, as only libraries of sufficient diversity of insert size and therefore sequence would be sequenced *en masse*. As each clone was uniquely identified by its coordinates in the 96 well plate, information on insert size was combined with paired end sequencing reads to assist in the assembly and validation of the sequence (section 3.5.5). 33-I19 *EcoRI*, *PstI* and *Sau3AI* sub-clone libraries and the targeted 5-F14 Sau3AI library were found to contain a diverse range of insert sizes (data not shown) and were therefore subjected to DNA sequencing.

3.5.4 Sub-clone library sequencing

Ninety six well plasmid preparations of each selected sub-clone library (section 3.5.2) were prepared (section 2.2.8) and sequenced using plasmid vector derived primers. Evaluation of each library (figure 3.11) demonstrated that there was both a low level of *E. coli* contamination (1 % in total) and as expected, representation of the Lawrist4 vector (the cosmid vector of 33-I19). The 1% *E. coli* contamination was acceptable and lower than expected (H. Davidson, Medical Genetics Section, University of Edinburgh, personal communication).

Approximately 5% of sequencing reactions were considered to have failed. This failure rate includes contaminated clones (more than one clone in a single sequencing reaction), plasmid vectors that contained no insert, instances where clones failed to grow and mechanical failure in the setting up of plasmid preparation and sequencing reactions.

The exceptional quality of the *PstI* digest derived library is evident from the total absence of *E.coli* contamination, the low proportion of Lawrist4 represented and the 100% success rate in sequencing this library (figure 3.11). The reduced representation of Lawrist4 can in part be considered an artefact of the way analysis was performed as any sequence with homology to both *Fugu* and Lawrist4 (*i.e.* the insert-vector junction) was counted as "Fugu" as it contains both *Fugu* sequence and positional information. Of the remaining three *PstI* fragments that could be contributed to by Lawrist4 two would be removed by the size selection step (section 3.5.2), and the remaining fragment would contain the origin of replication and so may be difficult to clone into a vector with its own origin of replication. The unprecedented success in sequencing this library and absence of *E. coli* contamination are inexplicable, but welcome phenomena.

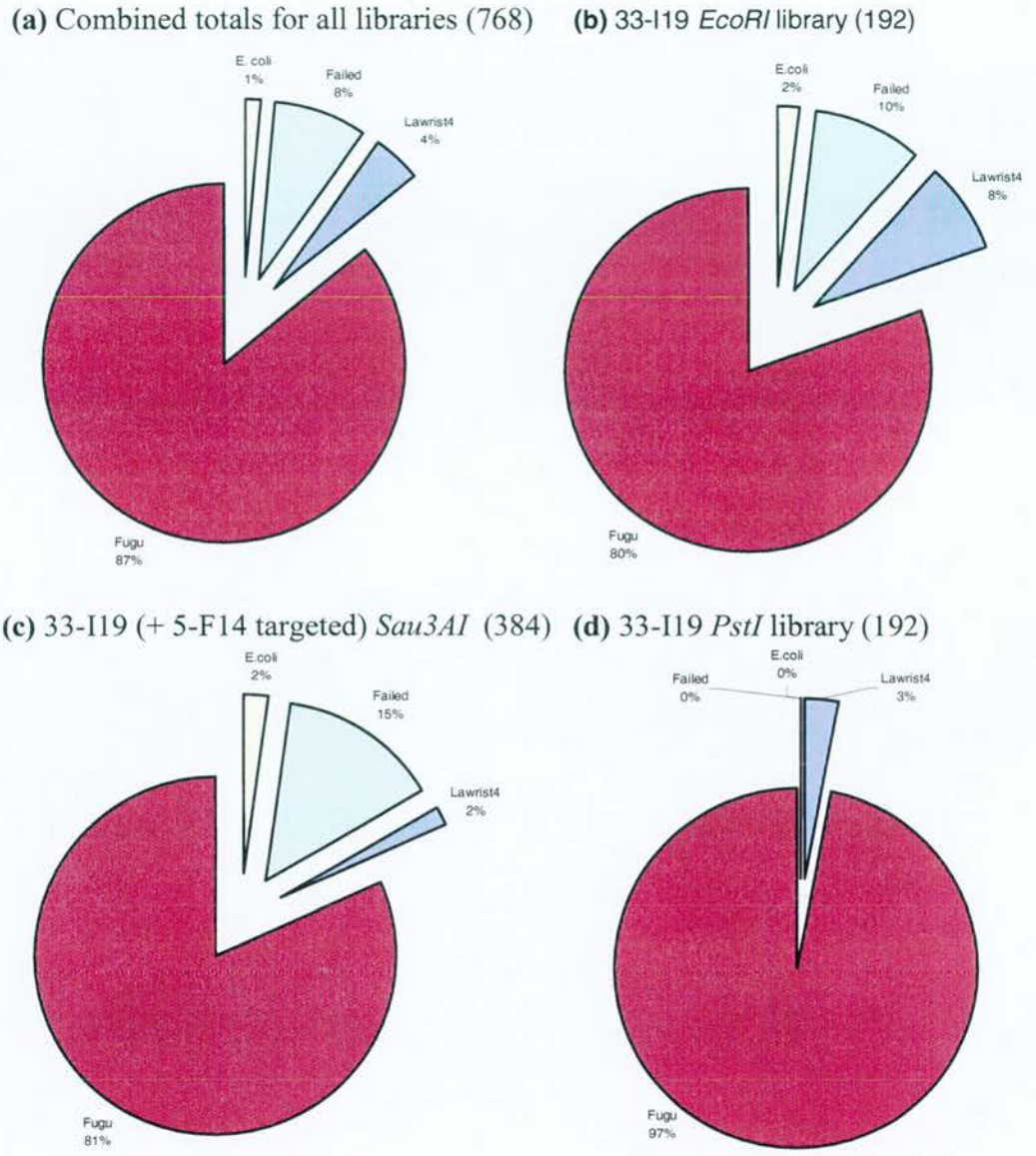


Figure 3.11; Analysis of sub-clone library sequencing. Statistics were calculated from unclipped (vector and poor quality sequence not removed) sequence by aligning (cross_match, section 2.11.2) to the finished genome of *E. coil* K12 (Blattner *et al.*, 1997), the Lawrist4 cosmid vector (section 2.2.4) or the assembled *Fugu* contig sequence (section 3.5.6). The category "failed" included sequencing reactions that failed, vectors that religated without incorporating an insert and mixed population clones. When a sequence contained both Lawrist4 and *Fugu* sequence it was categorised as *Fugu* for this analysis. Numbers in parentheses indicate the number of sequencing reactions per library. **(a)** Combined results

for all three libraries sequenced in full. **(b)** The 33-I19 *EcoRI* digest derived library. **(c)** The 33-I19 and 5-F14 targeted *Sau3AI* digest library. **(d)** The 33-I19 *PstI* digest derived library.

3.5.5 Sequence assembly

Chromatograms were interpreted and sequence assembled using the PhredPhrap software (Ewing *et al.,* 1998; Ewing and Green,1998) with modifications as described in section 2.11.2. During the PhredPhrap assembly process, sequences were screened against the Lawrist4 and pBluescript II SK- vector sequences, and matching sequence were masked prior to assembly. Sequences matching the *E. coli* genome were detected and removed after the assembly process.

Sequence contigs were manually edited through the Consed interface (section 2.11.2). Manual editing was required to overlap sequencing reads that were clearly related, but were not assembled into a single contig because of discrepancies in repeat length. CA repeats and tandem repeat arrays (figure 3.12) were particularly problematic.

Paired end sequences of the sub-clone libraries were assembled into 48 contigs (after filtering for *E. coli*) representing 33,785 base pairs of non-redundant sequence. Alignment to the completed reference sequence (section 3.5.6) demonstrated that this draft sequence represented 75.2% of the 33-I19 and 5-F14 cosmid insert sequences (figure 3.12).

3.5.6 Sequence finishing

Contigs from the draft sequence (figure 3.12) were extended by sequence walking (section 3.3.4). Oligonucleotides were designed optimally 80 nucleotides from the end of good quality sequence (phrap score of more than 20) using the primer picking tool within consed (section 2.11.2). An initial 94 oligonucleotides were synthesised for sequence walking, a further 28 oligonucleotides were used to complete the sequencing. Sequencing on sub-clone templates typically gave good quality sequence traces for 750 to 850 bp, whereas the same primers would only produced good quality traces of 400 to 600bp on cosmid templates. For this reason, the use of paired

end reads to select a sub-clone template for a particular sequencing reaction was crucial for the efficient finishing of the sequence.

The general lack of interspersed repetitive elements in the *Fugu* genome simplified the sequence assembly. However, tandem repeats appeared to be unusually abundant in this region of the *Fugu* genome (section 5.3.1). For the majority of these repeat arrays, the entire length of the tandem array was shorter than a single sequencing read so did not pose a problem for sequence assembly or finishing. The unusual interrupted pattern of the repeat shown in figures 3.12 was not well assembled by Phrap (section 2.11.2). This missassembly resulted in the failure to join other overlapping regions of the contig. Paired end reads, knowledge of sub-clone insert sizes and single nucleotide discrepancies between copies of the tandem repeat were used to manually assemble the sequence over this repeat region.

A complete sequence contig of 44,905 bp was assembled though sequence walking and manual resolution of Phrap missassemblies as discussed above. The whole contig comprising the combined inserts of cosmid clones 33-I19 and 5-F14 was covered by high quality data as defined by a Phrap score of ≥30 (Ewing & Green, 1998), except for a 661 bp region of low complexity sequence (figure 3.12). This 661 bp region contains multiple stretches of consecutive G residues which dramatically reduces sequence quality over the region resulting in the low quality scores. Assembly over this region has been confirmed by restriction analysis (section 3.4.3) and sub-clone paired-end read data (data not shown).

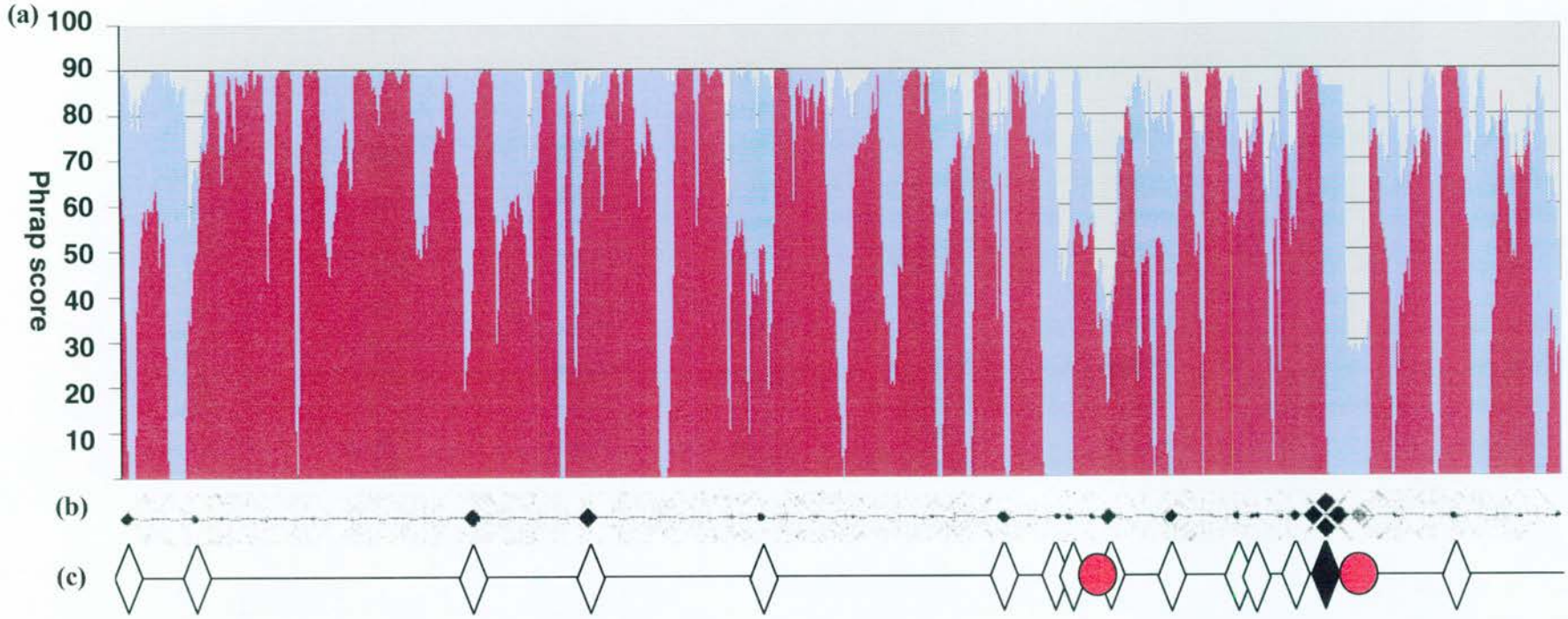


Figure 3.12; *Fugu* cosmid sequence assembly and finishing. **(a)** Sequence quality (Phrap score) of the 44,905 bp sequence contig. Phrap score is averaged for consecutive 50 bp windows over the length of the sequence. The red area plot indicates sequence generated by sub-clone library end sequencing, the blue area plot shows the quality of finished sequence over the contig. **(b)** Diagonal of the sequence contig plotted against itself (dotter, window size 17; section 2.11.2) highlighting regions of low complexity and tandem repeats within the sequence. **(c)** Sequence features identified by self versus self alignment, diamonds indicating tandem repeats of 3 or more nucleotides (mini-satellites) and red circles indicating mononucleotide, dinucleotide or other low complexity repeats. The internally disrupted tandem repeat responsible for sequence assembly problems (section 3.5.6) is indicated as a black diamond.

3.7 Discussion

Initial attempts at cloning *Fugu DISC1* genomic and cDNA clones directly through cross species hybridisation failed because of the poor sequence conservation of the *DISC1* coding sequence between *Fugu* and human. The progressively lower stringency hybridisation conditions used to detect distantly homologous sequences resulted in spurious hybridisation results that did not represent genuine homology between sequences. The identification of *TRAX* upstream of *DISC1* in humans provided a means of cloning a *Fugu* homologue of *DISC1* indirectly, using well conserved regions of the *TRAX* coding sequence as a cross species hybridisation probe.

A *Fugu* homologue of *TRAX* was readily isolated using the "double positive" strategy where two independent probes must hybridise the same clone prior to the time consuming validation of clones. The combination of cosmid restriction analysis, sequence sampling and cosmid end sequencing proved to be an efficient and rapid means of characterising the *Fugu* cosmid clones that had been isolated with human *TRAX* probes. This initial characterisation identified homologues of nidogen, *TM7SF1* and *DISC1* within the *Fugu* contig and demonstrated that the contig needed to be extended to clone the remainder of *Fugu DISC1*. By re-screening the *Fugu* cosmid library, the contig was extended by two cosmid clones. Further rounds of screening demonstrated that neither the cosmid library nor a *Fugu* genomic BAC library contained any clones that would extend the contig further downstream of *Fugu DISC1*.

The restriction digest based strategy for sequencing of cosmid clone 33-I19 and the targeted sequencing of 5-F14 was initially very efficient. The recognition of chimeric clones by virtue of them containing the particular restriction sites greatly simplified finishing of the sequence and sequence validation. However, the large number of walking steps required to finish the sequence illustrates the biased coverage introduced even when multiple restriction enzymes are used to construct libraries.

Not only was there a bias in sequence coverage, but also a bias of sequence quality was introduced (figure 3.12) that was compensated for in the finishing steps.

If such a sequencing project was to be undertaken in the future, a revised strategy of library production using both restriction digestion and sonication would be used. Such a combined strategy would provide a scaffold of sub-clones generated by restriction digestion whose chimeric status could be easily resolved, and a library with diverse inserts generated by sonication providing a means of generating unbiased shotgun sequence from the clones. This would reduce the need for sequence walking reactions that are both time consuming and costly compared to shotgun sequencing.

The high quality contiguous sequence generated from the *Fugu TRAX - DISC1* locus provides the basis for further comparative genomic, transcriptional and protein level analysis of this locus. These subjects are developed in subsequent chapters.

Chapter 4

The *TRAX – DISC1* region in vertebrates

4.1 Preface

DISC1 and *DISC2* were shown to be directly disrupted by the chromosome 1 breakpoint. The *TRAX* gene was serendipitously identified as a gene directly upstream of *DISC1* (Millar *et al.*, 2001b). However, chromosome translocations have been shown to exert long range position effects on proximal and distal genes (section 1.6.2). Therefore, an *in silico* analysis of the chromosome 1 breakpoint and flanking sequence was undertaken to characterise the wider genomic context of the region.

A localised assembly of clones and sequence fragments produced by the International Human Genome Sequencing Consortium (IHGSC) was carried out at two levels. The first level was to identify the genes adjacent to the breakpoint and determine their order and orientation. These genes would then be evaluated as positional and functional candidates for the t(1;11) phenotype. The second level of assembly was to produce contiguous sequence over a core "region of primary interest". This contiguous sequence would be used as the basis for molecular investigations of the human locus and comparative analysis with the contiguous *Fugu* sequence. The extent of this region of primary interest would depend in part upon the genes identified in the region and in part on the extent of conserved synteny with the orthologous *Fugu* locus.

The overall objective of using human to *Fugu* comparative genomic analysis to investigate the chromosome 1 breakpoint would be complemented by additional sequence from other organisms. As well as the potential benefit of using human to mouse comparative genomics, sequence of the mouse breakpoint locus was of interest for the future investigation of breakpoint genes in animal models. Therefore,

it was undertaken to obtain complete sequence for the mouse *TRAX – DISC1* locus. A novel strategy was devised utilising the assembled human genomic sequence and mouse genome informatic resources to identify mouse genomic clones covering the *TRAX – DISC1* locus.

4.2 The chromosome 1 breakpoint region

Using sequence and map resources of the human genome sequencing project (International Human Genome Sequencing Consortium, 2001) and the associated draft assemblies of the human genome, genes flanking the previously identified *TRAX* and *DISC1* genes were identified. Although the chromosome 1 breakpoint region was not well represented by finished sequence, BAC clone sequences could be confidently ordered based on sequence overlaps and FPC maps (section 2.11.1). Unlike the previous draft assemblies of the human genome, the Golden path genome assembly (http://genome.ucsc.edu/) based on the August 2001 "data freeze" was in good agreement with the manually curated ordering of sequences across this region. Using this August 2001 assembly as the reference sequence, BLASTX sequence similarity searching was used to identify the known genes and homologues of known genes within this genomic region. Their identities and associated evidence is summarised in table 4.1 Their relative positions and orientations are summarised in figure 4.1. Outside the region shown in figure 4.1, the arrangement of sequence fragments could not be confidently determined with the data available (data not shown).

The order and orientation of readily identifiable genes (table 4.1) in the *Fugu* clone contig is also summarised in figure 4.1 illustrating a boundary of conserved synteny with the *EGLN1*, *TRAX* and *DISC1* genes conserved in order and orientation between both humans and *Fugu*.

Gene[a]	Similar sequences[b]		Evalue[c]	Score[d]
	Accession	**Description**		
GNPAT	Q9NTV2	GNPAT (Human)*	9e-23	117
	Q9ES71	GNPAT (Rat)	7e-13	85
	P98192	GNPAT (Mouse)	7e-13	85
EXO84	O54924	EXO84 (Rat)	0.0	1280
	Q9NF17	CG6095 (*D. melanogaster*)	2e-53	219
	Q9VBI4	y105e8b.d (*C. elegans*)	9e-32	147
DJ876B10.3	Q9NTV0	Novel hypothetical protein (Human)*	1e-177	432
	Q22557	T19B10.6 (*C. elegans*)	2e-11	118
	Q9VXU6	CG9203 (*D. melanogaster*)	1e-19	102
EGLN1	Q99MI0	Cell growth regulator factor (Human)	4e-29	138
	Q62630	SM-20 (Rat)	2e-21	112
	Q9U4H6	Egg laying-9 (*C. elegans*)	7e-8	68
TRAX	Q99598	Translin associated protein X (Human)*	1e-64	256
	Q9JHB5	TRAX (Rat)	1e-64	256
	Q9VF77	CG5063 (*D. melanogaster*)	1e-14	90
DISC1	Q9NRI5	Disrupted in schizophrenia 1 (Human)*	1e-137	496
KIAA1389	Q9P2F8	KIAA1389 (Human)*	6e-55	222
	O95321	Putative GAP protein alpha (Human)	2e-10	75
	O43166	KIAA0545 (Human)	1e-5	59
KIAA1383	Q9P2G4	KIAA1383 (Human)*	0.0	1544
	Q9CUD6	4933403G14RIK (Mouse)	1e-159	562
Q9BSD7	Q9BSD7	Unnamed hypothetical protein (Human)*	2e-82	305
	Q9CQA9	2310079N02RIK protein (Mouse)	1e-66	253
	Q9SNJ0	Conserved hypothetical protein (Rice)	1e-17	90
PCNXL2	O43162	KIAA0435 (Human)*	4e-87	330
	O94897	KIAA0805 (Human)	2e-33	152
	P18490	Pecanex (*D. melanogaster*)	6e-31	143

Table 4.1; Protein coding genes predicted in confidently ordered human genomic sequence over the chromosome 1 breakpoint. Sequence masked for interspersed repetitive elements and low complexity repeats (RepeatMasker, section 2.11.2) was searched against the SPTR database (section 2.11.1) using BLASTX (section 2.11.2). Redundant sequence matches, sequence similarity representing processed pseudogenes (section 5.2) and similarity representing low compositional complexity were manually filtered out. The three highest scoring (BLASTX bit score) matches are summarised for each putative gene. **(a)** The gene symbol for each putative gene. Human genome organisation (HUGO) approved symbols were used where possible. The HUGO approved *EGLN1* gene symbol was assigned based

on work presented in chapter 9 of this thesis. **(b)** Non-redundant and filtered sequence similarity. Accession numbers refer to the SPTR database. The organism name for each sequence is indicated in parentheses. Asterisks indicate that the aligned sequence is >99% identical to the translated genomic sequence, and is therefore considered to represent the same gene. **(c)** BLASTX E-value for the reported alignment. **(d)** BLASTX bit score for the reported alignment.

4.2.1 The primary region of interest

As stated in the aims of this project (section 1.8), the primary region of interest for investigation is the chromosome 1 breakpoint, genes directly disrupted by the breakpoint and other genes in close proximity. The conserved synteny between human and *Fugu* over the breakpoint region extends upstream of *DISC1* to include the *EGLN1* and *TRAX* genes (figure 4.1). This upstream boundary of conserved synteny provided a convenient boundary to limit the primary region of interest. Downstream of *DISC1* in humans there is a large intergenic distance of up to 1 Mb before the next identifiable gene is encountered and there was no evidence for a homologue of known genes in the 10 kb of sequence downstream of *DISC1* in *Fugu*. As the *Fugu* sequence could not be extended further downstream of *DISC1* (section 3.4.3) the extent of *Fugu* sequence coverage downstream of *DISC1* provided a second convenient boundary to the region of interest. In humans this boundary of interest encompasses the *EGLN1*, *TRAX* and *DISC1* genes as well as the *DISC2* transcript. Each of these are investigated in subsequent chapters.

4.2.2 Wider genomic organisation

The genes defined as within the region of primary interest (figure 4.1; section 4.2.1) are intrinsically the strongest positional candidates for involvement in the t(1;11) phenotype, by virtue of their close proximity to or direct disruption by the breakpoint. However, as previously discussed, a chromosome translocation could affect the regulation of genes some distance from the breakpoint (section 1.6.2) or simply be in linkage disequlibrium with the true causative mutation (section 1.6.3). Within the confidently ordered and oriented draft human sequence, eleven genes could be identified including *EGLN1*, *TRAX*, *DISC1* and *DISC2*. The other identified genes were *GNPAT*, *EXO84*, *KIAA1389*, *KIAA1383*, *PCNXL2* and two hypothetical

genes that have no other name than database accession numbers *DJ876B10.3* and *Q9BSD7* (figure 4.1). Each of these genes flanking the region of interest are discussed briefly and their potential as candidates for a role in the t(1;11) phenotype discussed in section 4.5.

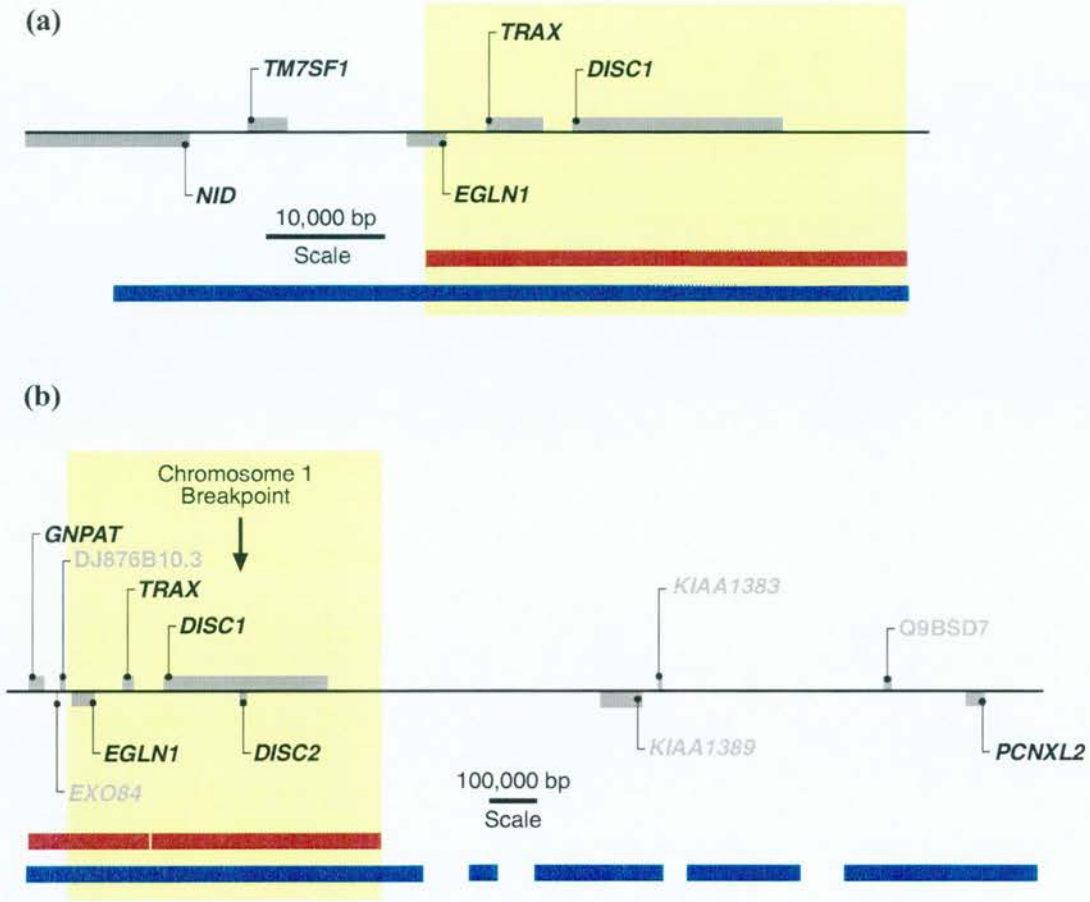


Figure 4.1; The wider gene organisation of the chromosome 1 breakpoint region. In both panels grey boxes indicate the extent of the transcription units of the identified genes. Those shown above the horizontal black line (representing genomic DNA) are transcribed left to right as shown on the figure, those shown below the horizontal black line are transcribed right to left. Yellow background indicates the "region of interest" as defined in section 4.2.1. **(a)** The *Fugu* genomic region orthologous to the chromosome 1 breakpoint locus. The solid brown line indicates the extent of contiguous sequence generated from the *Fugu* clone contig. The Blue line indicates the extent of the clone contig across which sequence sampling was carried out to identify genes. **(b)** The human chromosome 1 breakpoint locus. The brown line with a single gap indicates the extent of assembled human genomic sequence from the hybrid assembly (section 4.3) with the small gap of approximately known size indicated by the single gap. Blue lines indicate the contiguous coverage by clones that are in the process of being sequenced according to the August 2001 Golden Path genome assembly (section 2.11.1). The gene order shown has also been validated by the manually curated ordering and orientation of clone fragments and fluorescent *in situ* hybridisation (section 4.2.9 and data not shown).

4.2.3 *GNPAT*

The glyceronphosphate O-acyltransferase protein encoded by *GNPAT* is localised exclusively to the peroxisome where it is involved in biosynthesis of ether phospholipids (Thai *et al.,* 1997; OMIM: 602744). A partial deficiency of GNPAT enzyme function (as measured by biochemical assay) results in the syndromic phenotype of rhizomelic chondrodysplasia punctata (RCDP, OMIM: 215100). The RCDP phenotype is characterised by stippled foci of calcification in hyaline cartilage. Other symptoms of RCDP include coronal vertebral clefting, dwarfing, joint contractures cataract, ichthyosis and mental retardation (OMIM: 215100). In addition to mutations in the *GNPAT* gene, RCDP is also known to result from mutations in a range of genes whose protein products are peroxisome specific receptors or biogenic enzymes (Distel *et al.,* 1996; OMIM: 601539 for review). There have not been previous reports of an association of RCDP and mental illness, although the severe mental retardation associated with RCDP could make a diagnosis of mental illness problematic.

4.2.4 *EXO84*

The human *EXO84* gene has not previously been described, however, the single predicted exon annotated as human *EXO84* (figure 4.1) is supported by multiple ESTs and shows 92% identity in translation to the described rat *EXO84* gene (table 4.1). The EXO84 protein is well conserved in evolution with homologous proteins in budding yeast (Guo *et al.,* 1999) and mammals (Kee *et al.,* 1997). In both yeast and mammals, EXO84 has been demonstrated to be a component of the exocyst complex that is critical in mediating the cellular mechanism of exocytosis. (Guo *et al.,* 1999; Kee *et al.,* 1997). In budding yeast, null mutants of *Exo84* were not viable suggesting that this is an essential gene for eukaryotic cells (Guo *et al.,* 1999). Within yeast, Exo84 protein has also been found to be involved in pre-mRNA splicing, appearing to act as an enhancer of constitutive splicing (Awasthi *et al.,* 2001). Awasthi *et al.,* 2001 also confirmed the interaction of Exo84 with the Snp1p splicosome component by two-hybrid assay and co-immunoprecipitation.

4.2.5 *DJ876B10.3*

This hypothetical gene was supported by EST sequences (Unigene cluster Hs.23971) and represented by the ENSEMBL 1.2 predicted gene ENSG00000010072 (section 2.11.1). There were no known domains within this predicted protein and no indications of potential function from characterised homologous sequences. The predicted protein product of this gene is 65% identical to the uncharacterised *C. elegans* protein T19B10.6.

4.2.6 *KIAA1389*

The full length cDNA represents the full open reading frame of this predicted gene. Although not previously characterised, the *KIAA1389* gene contained several recognisable protein domains including the Rap_GAP (Pfam: PF02145) and PDZ (Pfam: PF00595) domains. The Rap_GAP domain is characteristic of the GTPase activating protein family of proteins that induce 'activated' G-proteins to hydrolyse bound GTP to GDP, returning the G-protein to the 'inactive' state (InterPro: IPR000331). This is an important general mechanism of intracellular signalling.

4.2.7 *KIAA1383*

The *KIAA1383* gene was represented by the full length cDNA sequence EMBL: AL451083 and Unigene cluster Hs.160373. The *KIAA1383* gene has not been previously characterised and has no informative homology with proteins of known function.

4.2.8 Q9BSD7

This hypothetical protein represented by the full length cDNA clone EMBL: BC005102 and multiple ESTs (Unigene cluster Hs.16034). The predicted 190 amino acid protein product had significant (E-value=0.00022) similarity to the AAA protein domain (SMART, section 2.11.2). The AAA domain is an ancient and highly conserved ATPase domain associated with a wide range of biological functions but typically involving a chaperone like mechanism (INTERPRO: IPR0011939).

4.2.9 *PCNXL2*

PCNXL2 is one of the two human homologues of the *Drosophila* protein *Pecanex*. The *Drosophila Pecanex* mutant is one of the "classic" maternal-effect neurogenic loci. In the absence of sufficient pecanex gene product, *Drosophila* embryos develop with severe hyperneuralisation similar to that observed in *Notch* mutant embryos (LaBonne *et al.,* 1989). *Pecanex* is transcribed both maternally and zygotically and continues to be neurologically expressed through all developmental and adult stages (LaBonne *et al.,* 1989). Fluctuating levels of *Pecanex* expression in *Drosophila* during development point to a role in patterning of the imaginal as well as the embryonic nervous system. The continued, albeit low level of *Pecanex* expression in the adult nervous system (LaBonne *et al.,* 1989) suggests a potential neural role for *Pecanex* in addition to patterning.

In mammals a *Pecanex* homologue was first described as a partial cDNA from rat (Gilbert *et al.,* 1992). Further work characterising the two mammalian homologues of *Pecanex* evident in the draft human genome (also represented by the full length cDNA clones AB023212 and AB018348) has not been reported although database submissions and reservation of the *PCNXL1* and *PCNXL2* gene names suggest a manuscript is in preparation.

Prior to the availability of ordered and oriented sequence from the human genome sequencing project, the *PCNXL2* gene was identified as being in the general region of the chromosome 1 breakpoint. This mapping was based on the STS content of human genomic sequence containing fragments of the *PCNXL2* gene. The obvious relevance of a gene crucially involved in neural patterning and development that was potentially within the region of interest lead to a brief investigation into the position of *PCNXL2*. A small contig of BACs was constructed over the *PCNXL2* locus (not shown). A BAC containing the *PCNXL2* gene was used in two-colour fluorescent *in situ* hybridisation with a breakpoint crossing BAC (figure 4.2). Results of the *in situ* hybridisation indicated that the *PCNXL2* locus was 1.5 to 3 megabases telomeric to the chromosome 1 breakpoint a result that validates the assembly of genomic sequence shown in figure 4.1.

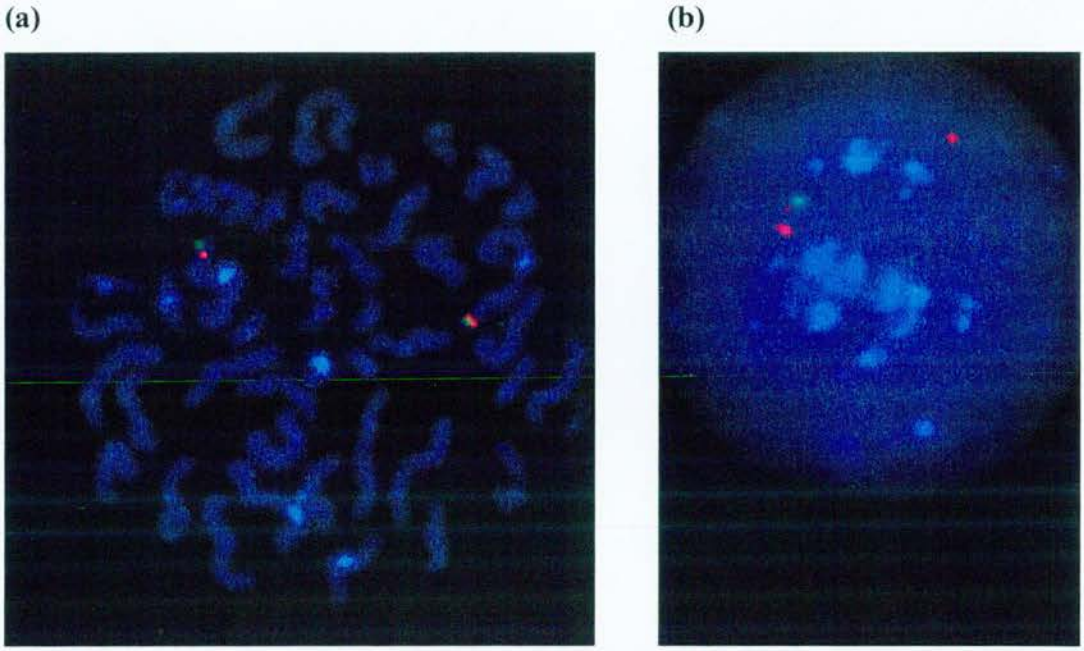


Figure 4.2; Fluorescent *in situ* hybridisation of a breakpoint crossing PAC and a *PCNXL2* containing PAC clone. The chromosome 1 breakpoint crossing PAC is labelled with a red fluorescent dye and the *PCNXL2* a green fluorescent dye. **(a)** Representative hybridisation of a metaphase spread. The only detected signal is on the distal end of chromosome 1. The *PCNXL2* containing PAC hybridised telomeric to the breakpoint. **(b)** Representative hybridisation of an interphase nucleus. Based on interphase nuclear signals, the *PCNXL2* gene was estimated to be 1.5 to 3 Mb telomeric to the chromosome 1 breakpoint.

4.3 Assembly of human sequence

4.3.1 Two draft assemblies of the human genome

Two landmark papers published in February 2001 reported the completion and assembly of "working drafts" of the human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.,* 2001). The two strategies for genome sequencing and assembly are summarised in figure 4.3.

The International Human Genome Sequencing Consortium (IHGSC) strategy was based on constructing contiguous clone maps over the euchromatic region of each chromosome and selecting a minimal tiling path of clones to sequence (typically BAC clones). The selected clones were then broken into smaller fragments and the smaller fragments sequenced from both ends, a strategy described as "shotgun sequencing". The shotgun sequences from each BAC clone were then overlapped to produce sequence contigs, the sequence contigs were ordered and oriented using the paired end reads of shotgun fragments. A clone in this status is referred to as a phase 0 or phase 1 draft. Subsequently draft clones were finished by performing additional sequencing reactions and validated by restriction analysis and / or PCR reactions. A finished clone will typically have every base of the clone insert sequence covered by at least five sequencing reactions in each direction. A finished clone is expected to have less than one error per 1,000 bp (IHGSC, 2001). Although a draft assembly has been released, the hierarchical sequencing is ongoing with new clones being sequenced and clones being finished. As of the 1st of April 2001, it was estimated that 40% of the genome was represented by finished sequence and a further 50% is represented by draft sequence (J. Kent, personal communication). Sequences of clones are released by the IHGSC in phase 0 status and are updated as the sequence is improved.

The draft human genome sequence reported by Venter *et al.,* (2001) was based on a whole genome shotgun sequencing strategy where the entire human genome was broken into fragments of approximately known sizes (2 kb, 10 kb and 50 kb

libraries). The ends of each fragment were sequenced, with information relating to mate-pairs (the opposite end sequences of a single fragment) and expected fragment size retained for the assembly process. The assembly process used by Venter *et al.,* (2001) incorporated proprietary whole genome shotgun sequence generated by Celera Genomics and sequence from the IHGSC. The IHGSC sequence incorporated into the assemblies of Venter *et al.,* (2001) was "shredded", i.e. broken up into 550 bp fragments. The associated ordering and orientation information of these fragments was not retained for use in the assembly (Venter *et al.,* 2001).

While the publicly available assembly of the human genome from Celera Genomics (Venter *et al.,* 2001) is not updated with new sequence data, both the National Center for Biotechnology Information (NCBI) and the University of Santa Cruz (UCSC) are continuing to incorporate newly generated sequence into refined assemblies of a reference human genome (Taylor, 2001a for review). Because these assemblies attempt to use as much information as is available to produce the best possible assembly there is little data that can be used to independently validate the quality, coverage and accuracy of the assemblies. However, using independent data sets or direct comparison methods several assessments of the assemblies have been made (Aache *et al.,* 2001; Oliver *et al.,* 2001; Wright *et al.,* 2001; Semple *et al.,* 2002). These investigations have found dramatic discrepancies between assemblies and misassemblies at whole chromosome and small scale levels. Of particular note is the comparison of draft genome assemblies by Semple *et al.,* (2002) that focuses on a region of the genome that is of a similar scale to the chromosome 1 breakpoint region investigated in this work. Within the 5.8 Mb region analysed by Semple *et al.,* the draft assemblies were found to have between 2 and 5 miss assemblies per megabase, as well as failing to include sequence known to be from the region and the erroneous inclusion of sequence from elsewhere in the genome.

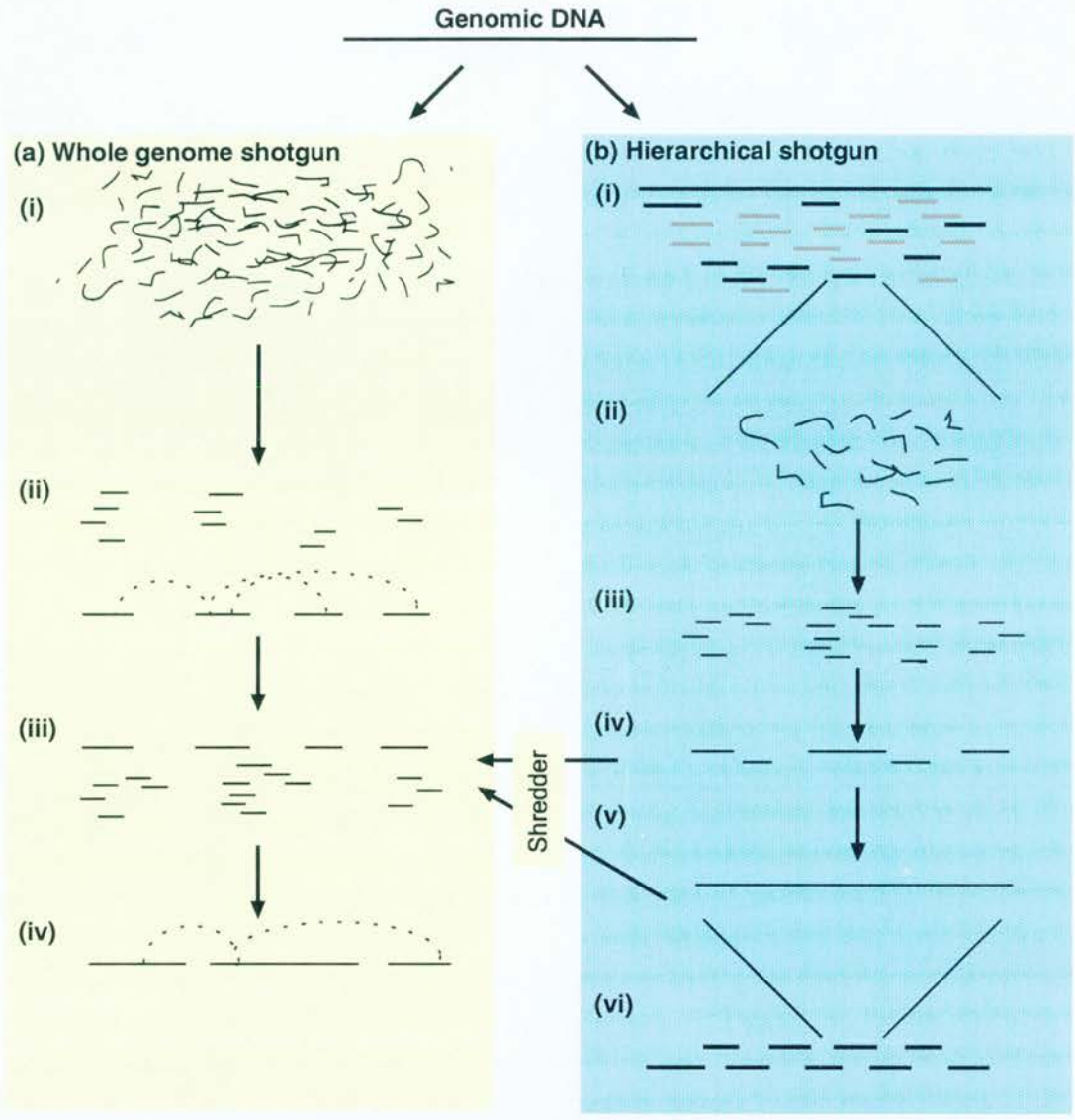


Figure 4.3; Two drafts of the human genome. **(a)** The whole genome shotgun strategy used by Venter *et al.*, (2001). ***(i)*** Total genomic DNA was fragmented (physical shearing forces) into fragments of a pre-determined size range. ***(ii)*** Fragments were cloned into plasmid vectors and each end of the inserts sequenced. The known inverted relative orientation of opposite end sequences, the approximately known length of intervening sequence as well as sequence overlaps with other fragments were used to assemble, order and orient sequences. ***(iii)*** The assembly process also incorporated 550 bp sequence fragments from the IHGSC project to extend contigs and fill gaps between contigs (indicated by "shredder" in the diagram). ***(iv)*** A final assembly with blocks of contiguous sequence ordered and orientated based on opposite end sequence information. Gaps between sequence contigs were of an approximately known size, these are refered to as sequence scaffolds. Scaffolds themselves were ordered, oriented and mapped to chromosomes using STS marker

information. **(b)** The hierarchical shotgun sequencing approach used by the IHGSC, (2001). ***(i)*** A contig of large insert clones (100-250 kb) across the whole genome was constructed using STS marker hybridisation and fingerprint analysis among other techniques. A path of minimally overlapping clones (minimal tiling path) was selected across a region (indicated by horizontal black lines). ***(ii - iii)*** Individual selected cones were shotgun sequenced. ***(iv)*** Shotgun sequences were assembled incorporating paired end sequence data in a similar manner to Venter *et al.,* (2001). ***(v)*** Gaps between sequence contigs were filled by additional sequencing reactions "finishing" and the final sequence validated by restriction analysis. ***(vi)*** Sequence overlaps between finished clone sequences are then determined and a contiguous sequence between clones is derived.

4.3.2 A hybrid assembly of the chromosome 1 breakpoint region

Gaps and misassemblies in the available drafts of the human genome (section 4.3.1) would adversely affect the primary sequence annotation and subsequent comparative genomic analysis of the chromosome 1 breakpoint region. As both the Celera assembly and sequence from the IHGSC contained complementary ordering and orientation information as well as each possessing sequence not represented in the other data set it was considered likely that a hybrid assembly could be produced that was more accurate and had greater coverage than either data set alone. This strategy did not conflict with terms of the Celera licence agreement due to the small scale (less than 1 Mb) of the region being assembled and that the sequence for the assembly was not released, only the appropriate information to recreate it. Combining sequence as well as order and orientation information from both the Celera Genomics (Venter *et al.,* 2001) and the IHGSC whole genome sequencing project, a hybrid assembly of the region around the chromosome 1 breakpoint region was undertaken. The hybrid assembly strategy also included the integration of in-house data regarding marker and gene order.

IHGSC clones over the chromosome 1 breakpoint region were identified by sequence similarity searching (BLASTN, section 2.11.2) with human *DISC1* and *TRAX* cDNA sequences. Overlapping clones were identified by reference to the accession maps project (http://www.genome.wustl.edu/gsc/human/Mapping/). Fifteen IHGSC clones in a total of 186 fragments of contiguous sequence were identified as representing the region of interest (table 4.2). BLASTN similarity searching with human *TRAX* and

DISC1 cDNA sequences was also used to identify the Celera sequence scaffold GA_x8YCEG9, the first megabase of this 2.4 Mb scaffold representing the region of interest.

The first 1 Mb of scaffold GA_x8YCEG9 consisted of 73 ordered and oriented sequence contigs. For the purposes of producing a hybrid assembly, the scaffold was fragmented into each of its constituent contigs but ordering and orientation information was retained in the fragment naming scheme. The 186 IHGSC fragments were broken into separate sequences although the naming scheme retained information of fragment chaining (ordering and orientation information sometimes included in an IHGSC clone sequence accession) and vector end information (sequence fragments that represent the ends of an IHGSC clone insert). A significant feature of the assembly process used was to weight sequence fragments based on their expected accuracy. In the assembly described, finished IHGSC clones which were anticipated to be of very high quality (section 4.3.1) were given an artificial Phrap score (Ewing and Green, 1998) of 60 where as Celera fragments and unfinished IHGSC clone sequences were given Phrap scores of 15. These weighted scores were empirically determined to work well for this particular assembly. It may be possible to improve performance by further optimisation of these scores and the introduction of a sliding scale for differing levels of assembly.

An initial fully automated assembly of fragments was carried out using the Phrap.longreads (section 2.11.2) software. The initial assembly was then manually edited in the Consed interface (section 2.11.2). Short sequence overlaps (typically 30 to 79 bp overlap of 100% identity) not permitted by Phrap.longreads were allowed if they were consistent with Celera or IHGSC fragment ordering. All instances of discrepancies between reported chaining and the sequence assembly were individually investigated and resolved by taking a consensus of overlapping sequences. The basic logic used was that a sequence overlap dominated a report of chaining. Tests were also carried out to ensure that all sequence from an IHGSC clone was located within the assembly between the ends of the clone insert. One instance where this was found not to be the case (clone accession AL540284, see

figure 4.4) was subsequently resolved by the next update of that clone sequence. The offending misplaced sequence was removed from the clone indicating that it represented sequence contamination.

The end result of this assembly process was two blocks of contiguous sequence (455,334 bp and 331,910 bp) across the entire region of primary interest (section 4.2.1). The single gap between blocks of sequences was predicted by the Celera scaffold to be only 147 bp in length, a prediction that is consistent with the IHGSC clone that bridges the gap. The full coordinates of this hybrid genomic assembly are given in appendix I.ii. This assembled sequence forms the basis of subsequent analysis of the region and its comparison with the orthologous *Fugu* locus.

Accession number[a]	Version[b]	Fragments[c]	Centre[d]	Length[e]
AC011655	4	19	WIBR	175106
AC011945	4	42	WIBR	152011
AL117352	12	1*	SC	138056
AL136171*	17	1*	SC	133968
AL161743*	20	1*	SC	100170
AL353052	3	11	SC	151132
AL356745	6	26	SC	161749
AL358784	4	12	SC	202092
AL359543	5	8	SC	148291
AL359814	4	5	SC	138756
AL445200	1	16	SC	133054
AL445283	4	35	SC	142058
AL445524	19	5	SC	227526
AL450284	6	1*	SC	66208
AL589644	3	3	SC	168386

Table 4.2; IHGSC BAC clone sequences used in the hybrid sequence assembly. **(a)** EMBL accession numbers for clone sequences. Accession numbers marked with an asterisk indicate that the clones are PAC rather than BAC. **(b)** Version number refers to the sequence version number used in the assembly. **(c)** Fragments indicates the number of sequence contigs within the clone. Finished sequences are indicated with an asterisk. **(d)** Acknowledgement of the sequencing centre that generated the sequence. SC indicates the Sanger Centre, WIBR indicates the Whitehead Institute / MIT Center for Genome Research. **(e)** Total length of sequence in the entry (base pairs).

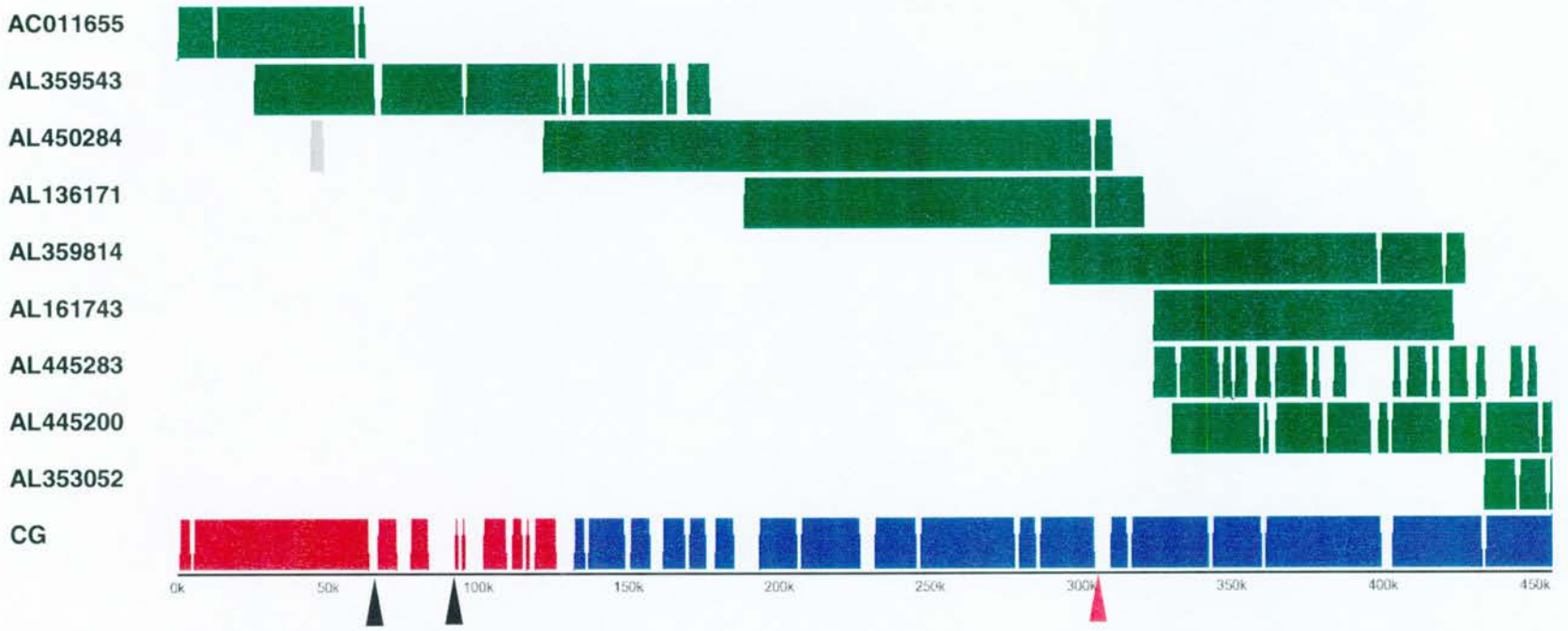


Figure 4.4; A hybrid assembly of human genomic sequence. Accession numbers for IHGSC large insert clone sequences are shown down the left of the figure and correspond with the horizontally aligned bars. GC indicates Celera Genomics human sequence scaffold GA_x8YCEG9. The horizontal axis of the graph represents the 455334 bp contiguous hybrid sequence assembly over the entire *DISC1* genomic region (section 4.3.2). The graph shows a modified Vista plot (section 2.11.2) of avid (section 2.11.2) alignments between the sequences used for the assembly and the final hybrid assembly sequence. Each coloured block represents contiguous sequence of >95% identity (>99% of alignments were >99% identical). Continued on the next page.

Gaps between blocks indicate a lack of sequence coverage. The block coloured light grey represents sequence that was erroneously included in the AL450284 sequence entry and was removed in the subsequent finished version of this sequence (version 6). Black triangles below the graph indicate regions that are not obviously covered by any of the aligned sequences. There are short sequence overlaps between blocks of AL359543 and the Celera sequence scaffold that are not readily visible at the resolution shown. The red triangle indicates a CT rich low complexity sequence of approximately 300 nucleotides in length that varied in length between all three sequences over the region. The sequence of AL359814 was arbitrarily chosen as the reference sequence over the region. Red and blue colours of the Celera sequence scaffold indicate two separate regions of the scaffold that were erroneously separated by the insertion of sequence resulting from a localised missassembly.

4.4 Identification of the mouse *TRAX – DISC1* region

4.4.1 Introduction

A collaboration of genome sequencing centres around the world is currently coordinated in an effort to map and sequence the whole mouse genome. The strategy is one of whole genome shotgun to 4x - 6x coverage combined with a whole genome FPC map, BAC end sequences and BAC clone by BAC clone finishing. The eventual aim is a high quality, finished euchromatic portion of the genome (J. Rogers, personal communication). The standard reference libraries for the public mouse sequencing project are the BAC libraries RPCI-23 and RPCI-24 (female and male respectively) derived from the C57BL/6J mouse strain.

The UK Mouse Sequencing Consortium, working within the framework of the Public Mouse Sequencing Consortium, is prioritising the sequencing and finishing of regions of the mouse genome that are of particular biological importance. Through collaboration with the UK Mouse Sequencing Consortium, sequencing of the mouse *TRAX – DISC1* genomic region was undertaken.

Cloning of mouse *TRAX* (R. Devon and section 8.3) and the isolation of genomic PAC clones containing *TRAX* and the first eight exons of *DISC1* (R. Devon) led to the demonstration that *TRAX* and *DISC1* were in the same order and orientation as in

human and *Fugu*. These previously identified mouse genomic clones did not cover the entirety of the region to be sequenced, were not from the reference libraries and were consequently not appropriate to be sequenced. It was therefore necessary to identify a minimum tiling path of clones from the RPCI reference libraries that could be sequenced.

Although the libraries RPCI-23 and RPCI-24 were available as gridded filters for screening by hybridisation, a significant amount of work would be necessary to identify candidate clones, assemble the contigs and determine a minimum tiling path of clones prior to submitting them for sequencing. As an alternative strategy, the existing RPCI-23 / RPCI-24 FPC maps (section 2.11.1) and BAC end sequences (obtained from the EMBL database, section 2.11.1) were used to identify contigs containing the *TRAX – DISC1* region and select a minimum tiling path of clones to sequence across the region. The underlying rationale of this approach was based on the expectation that the majority of matches between non-repetitive human genomic sequence and the orthologous mouse BAC end sequence would be reciprocally, the best match between the whole human genome and all mouse BAC end sequences.

4.4.2 Mouse FPC contigs

Human assembled sequence (section 4.3.2) was masked for known repeats (RepeatMasker, section 2.11.2) and searched against the mouse BAC end sequence database (MBENDS, section 2.11.1) using BLASTN. All matches with an E-value of less than 0.01 were considered for further analysis, an arbitary and liberal cut off. Eight BAC end sequences matched this criteria (table 4.3).

BAC end sequences from clones RPCI-23-393P20 and RPCI-23-392E17 from ctg2550 were >99% identical to one another, consequently these clone pairs were not considered to be independent evidence that a contig ctg2550 was orthologous to the human TRAX-DISC1 region. The BAC end sequences from clones RPCI-24-244F19, RPCI-24-244G19 and RPCI-23-121A6 also shared a high degree of sequence identity and matched human sequence only within a processed pseudogene for SNRP-D2 these clones were consequently excluded from further analysis.

BAC end sequences for each of the remaining mouse FPC contigs (ctg516, ctg2550 and ctg6022) were identified from the EMBL database. Every available BAC end sequence was masked for repeats (RepeatMasker using the rodent repeat database, section 2.11.1) and searched back against the Golden Path human genome assembly (April 12, 2001 data freeze) using BLASTN. The purpose of this back validation was two fold. The first was to discover if the best match (highest BLASTN bit score) in mouse BAC end sequences also found the human query sequence to be the best match in the human draft genome. This principle of a reciprocal best match is an important validation that should dramatically reduce false matches due to low copy number repeats, segmental duplications and pseudogenes. The second reason for back validation was the extended range of Golden Path scaffolds relative to the curated human assembly, which have the potential to highlight approximate regions of synteny breaks and provide further evidence of homology between the region of interest and candidate FPC contigs.

Within the April 12th 2001 Golden Path assembly of the human genome, the human *TRAX - DISC1* region was contained within the 15 Mb scaffold ctg4256 (BLASTN, data not shown). Searching the 133 available mouse BAC ends from ctg2550 back against the Golden Path assembly identified 13 sequences with sequence similarity (BLASTN bit score range of 44 to 581) to human scaffold ctg4256. In 12 of these cases scaffold ctg4256 was the highest scoring match. However, all of the matches were to a region of the human contig located at least 8 Mb from the *TRAX – DISC1* locus. Therefore mouse FPC contig ctg2550 is syntenic to human chromosome 1, but not the *TRAX – DISC1* locus.

Of the 77 available BAC end sequences from the mouse FPC contig ctg516, 12 had detectable similarity to the human ctg4256 sequence. In 11 of these cases, alignment with ctg4256 represented the best scoring alignment out of all available human genomic sequence. Eleven of these matches were to human sequence located within the 2 Mb flanking the *TRAX – DISC1* locus. By the same criteria, mouse FPC contigs ctg6022 and ctg3436 also matched the *TRAX – DISC1* locus better than any other region of the human genome (figure 4.4).

As an additional round of contig investigation and validation, the cDNA sequences for mouse *TRAX* and *DISC1* (sections 8.3 and 6.4.3 respectively) as well as associated genomic sequence from the mouse whole genome shotgun sequencing project (section 6.4.3) were searched against the mouse BAC end sequence database using BLASTN. As this additional round of searching used mouse versus mouse sequence exact or nearly exact matches were expected. Therefore, only sequence identity >95% over >50 nt was considered for further analysis. The results of this analysis further supported the conclusion that FPC contigs ctg4256, ctg6022 and ctg3436 represented the mouse *TRAX – DISC1* locus (figure 4.4).

Specifically selecting the three identified mouse BAC contigs and reducing the stringency of FPC overlapping, the three contigs were all found to overlap in a manner consistent with the observed order of BAC end sequences and the homologous sequences in the human genome (figure 4.4). Alignment between FPC contigs and assembled human genomic sequence suggested that the mouse locus was compact compared to the human locus (figure 4.4). The apparent compaction of 880 kb of human genomic sequence into less than a BAC length (less than 300 kb) of mouse genomic sequence illustrated in figure 4.4 may be an artefact of misassembled human genomic sequence. The other two regions of compaction were over well curated human sequence assemblies (section 4.3.2 and data not shown) and are supported by accurately sized human clone contigs (K Millar and S Christie, personal communication). Similar reductions in the size of mouse loci relative to the orthologous human locus have also been reported for several other regions of mouse versus human comparison (Taylor, 2001a for review).

Clone name	Clone end	E-value of match	FPC contig location
RPCI-23-418L11	T7	2e-20	ctg516
RPCI-24-244F19	T7	4e-15	ctg1636
RPCI-24-244G19	T7	4e-15	ctg2709
RPCI-23-121A6	T7	2e-11	ctg2709
RPCI-23-393P20	T7	1e-9	ctg2550
RPCI-23-392E17	T7	1e-9	ctg2550
RPCI-23-141E24	T7	1e-9	ctg516
RPCI-23-236F19	T7	0.004	ctg6022

Table 4.3; Mouse BAC end sequences showing sequence similarity to the assembled human TRAX-DISC1 genomic sequence. FPC contig location refers to the mouse FPC contig number based on mouse map release 010418 that the clone is contained within (http://www.bcgsc.bc.ca/projects/mouse_mapping/).

4.4.3 Clone selection

Based on BAC end sequence homology and FPC clone overlaps, eleven clones were obtained for further validation. Clones were selected that would give several options for tiling paths dependant on the outcome of further validation (section 4.4.4). All selected clones were from the RPCI-23 library, they were: 144L1, 236F19, 12D24, 42A14, 538K3, 33O1, 14124, 127A6, 356I16, 9L16 and 418L11. See figure 4.4.

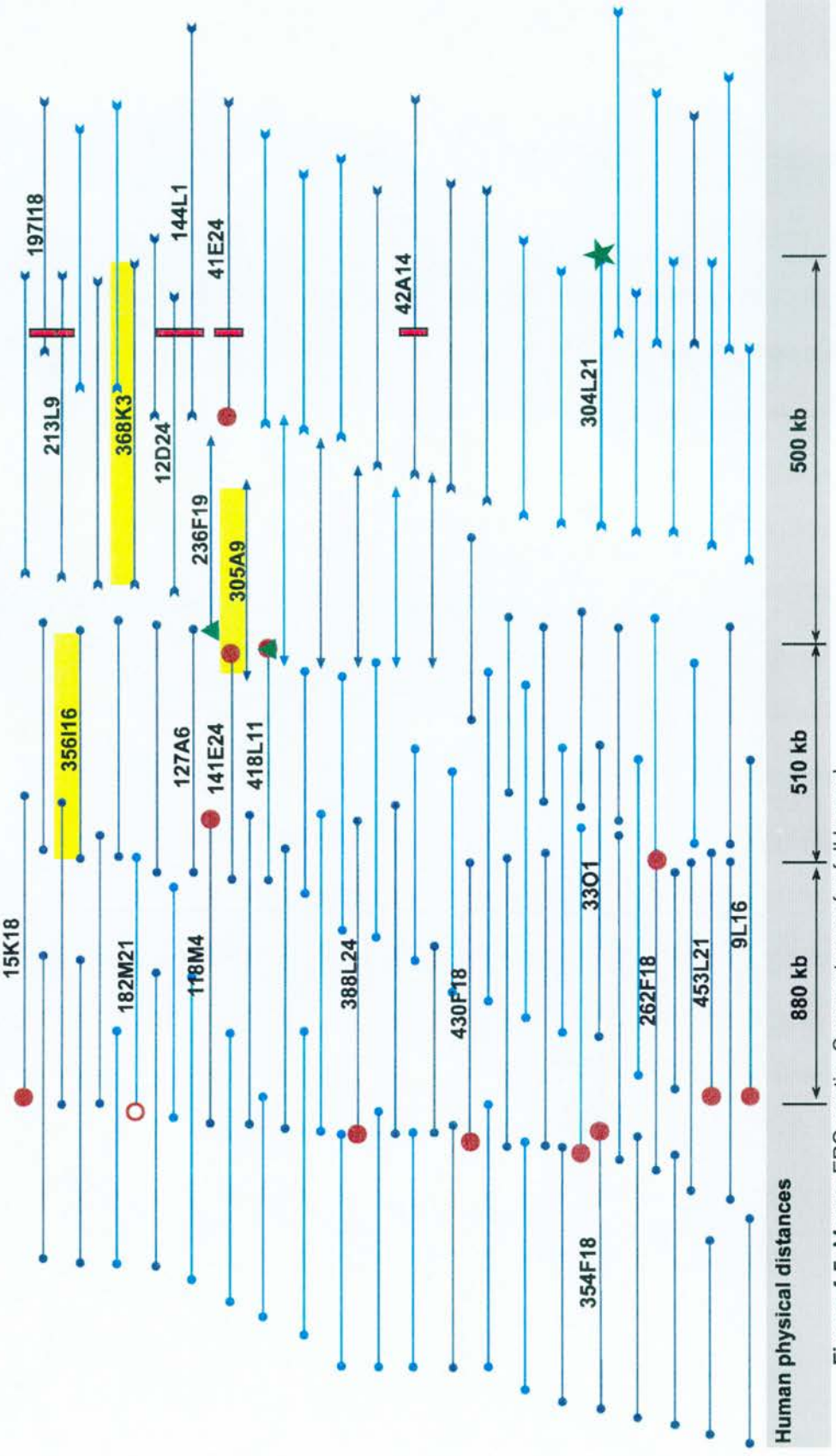


Figure 4.5; Mouse FPC contig. See next page for full legend.

Figure 4.5; Mouse FPC contig. Horizontal blue lines indicate mouse BAC clones. Dark blue indicates the RPCI-23 library clones and light blue the RPCI-24 library clones. All clones that were selected for evaluation and clones with identifiable homology to known human or mouse sequences are labelled with the clone name. The FPC contig shown is a merged contig derived from three separate FPC contigs (mouse map release 010418, section 2.11.1) Clones from each contig are distinguished by shapes at the ends of the horizontal line: ctg516 (small circles), ctg6022 (arrow heads), ctg2550 (arrow tails). The clones selected for sequencing are highlighted in yellow. Red circles at a clone end indicate BAC end sequences that when searched against the April 12 2001 data freeze human genome assembly (section 2.11.1) the highest scoring BLASTN match was to human contig ctg4256 in the 2 Mb flanking the *TRAX* and *DISC1* genes. The open red circle of clone 182M21 indicates that although this BAC end sequence did match the human sequence contig ctg4256, the highest scoring match was elsewhere in the human genome. Green triangles indicate BAC end sequences that showed >99% sequence identity in alignment with mouse *TRAX* cDNA sequence (section 8.3). The green star indicates BAC end sequence that was >99% identical in alignment with mouse sequence identified as from *DISC1* intron 9 by iterative sequence clustering (section 6.4.3). The red rectangle indicates BAC clones reported (FPC mouse map release 010418) to be positive by PCR or hybridisation to the mouse STS D8Mit42 that was identified to be in tight linkage with the mouse *TRAX* gene (Devon *et al.*, 2000). The relative orientation of BAC end sequences compared to the assembled human genomic sequence was used to infer the orientation of each mouse BAC clone and to use as cross species sequence "anchors" for the comparison of physical distance between the mouse contig and human sequence assembly (indicated in the grey bar below the mouse contig). There is some evidence (not shown) that the human assembly used for comparison contains a misassembly. This misassembly would cause the 880 kb physical size estimate of the human genome between anchors shown to be an over estimate. The estimates of 510 kb and 500 kb are consistent with the hybrid human sequence assembly (section 4.3.2) and the independent clone contig over the region (K Millar, personal communication).

4.4.4 Clone validation

Selected mouse BAC clones were validated by *HindIII* restriction digestion and comparison to the expected FPC restriction fragment profile (figure 4.5). As the fragments were visualised by agarose gel electrophoresis and transillumination rather than radio active or fluorescent labelling, only the larger bands were visible (figure 4.5). However, comparison of the restriction digest profiles with FPC profiles was sufficient to identify if the expected clone had been isolated (figure 4.5). The BAC

clone DNA was bound to nylon membranes and hybridised with oligonucleotides designed to exons of the mouse *DISC1* gene (section 6.4.3). These hybridisation experiments and PCR reactions (figure 4.6 and data not shown) confirmed that the mouse *TRAX – DISC1* locus had been identified. Based on FPC overlaps and hybridisation results, the BAC clones 365I16, 305A9 and 368K3 were submitted to the UK Mouse Genome Sequencing Consortium sequencing pipline (experimental validation of clone 305A9 was carried out by R. Moore, University of Edinburgh). Unfortunately the sequence of these clones will not be available until after the submission of this thesis. However, the complete open reading frames and genomic structures of mouse *TRAX* and *DISC1* have been independently obtained and analysed (section 8.3 and 6.4.3 respectively).

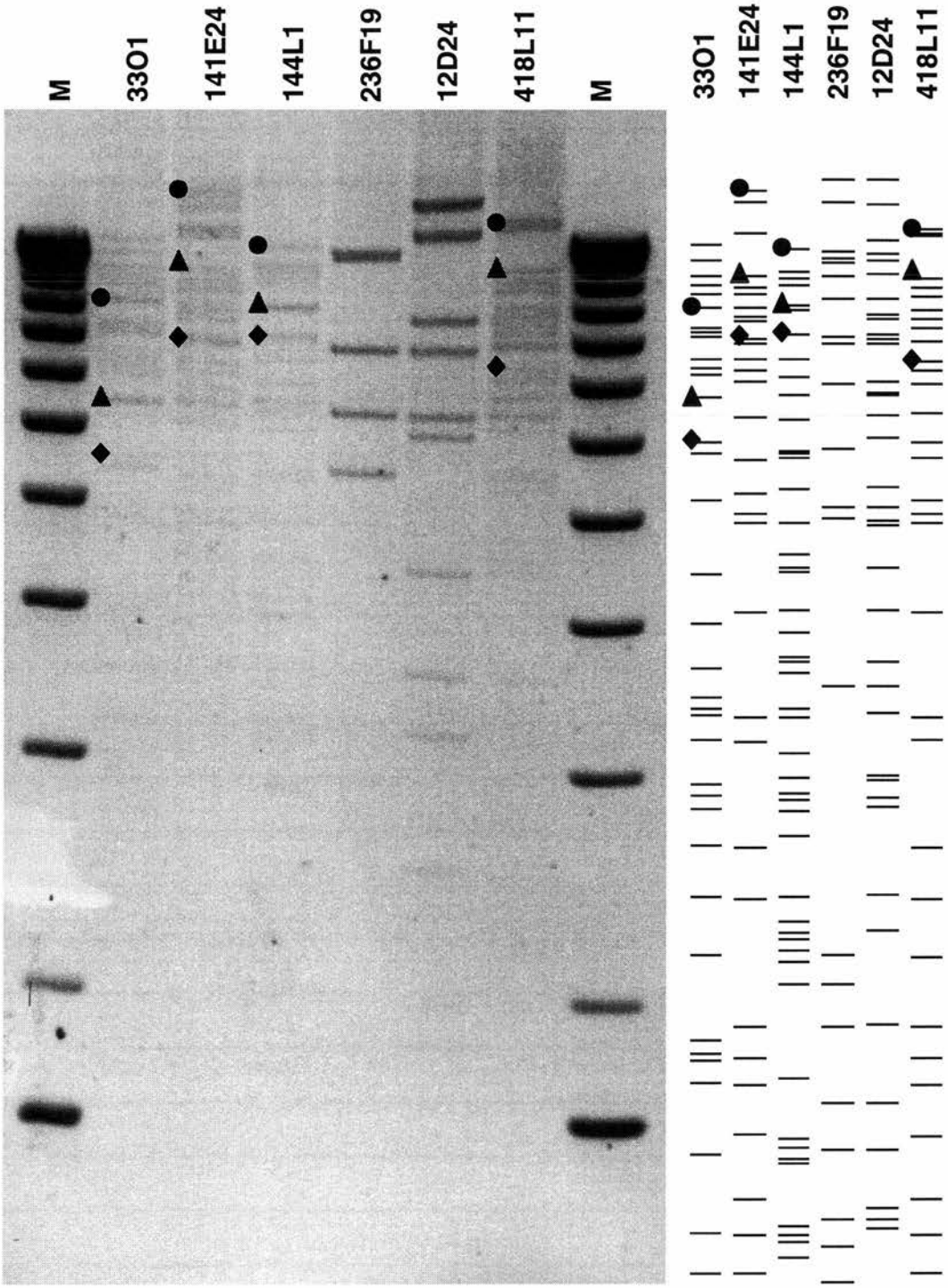


Figure 4.6; **(a)** Mouse BAC fingerprint analysis. Continued on the next page.

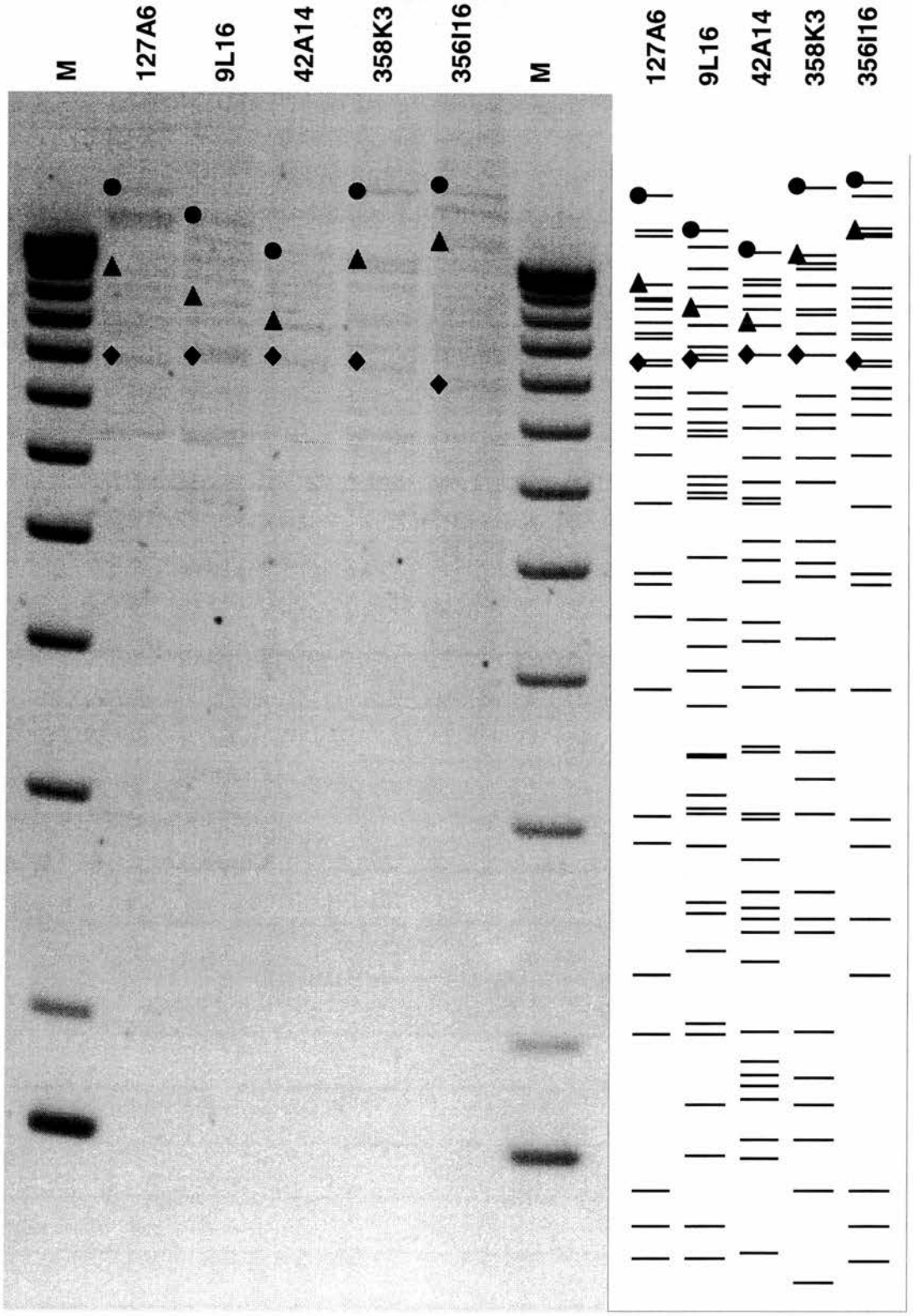


Figure 4.6; **(b)** Mouse BAC fingerprint analysis. Full legend on the next page.

Figure 4.5; Mouse BAC fingerprint analysis. *HindIII* digests of selected mouse BACs electrophoretically separated on 0.8% agarose gels and visualised by UV transillumination. Lanes marked 'M' indicate the 1 kb DNA molecular weight marker. To the right of the gels, the FPC fingerprints for the corresponding clones are shown. The fingerprint image was uniformly stretched to approximately align the bands with the agarose gel image. Black circles, triangles and diamonds indicate gel bands that are interpreted as representing the correspondingly annotated FPC band. **(a)** The *HindIII* restriction patterns of BAC clones 33O1, 141E24, 144L1 and 418L11 closely match those expected from FPC fingerprint profiles. Clones 236F19 and 12D24 do not match the expected FPC fingerprint profile and appear to have very small inserts. These may have been contaminated by other clones or represent internally deleted derivatives of the expected clones. **(b)** The HindIII restriction pattern of clones 127A6, 9L16, 42A14, 358K3 and 356I16 closely matches the expected FPC fingerprint profiles.

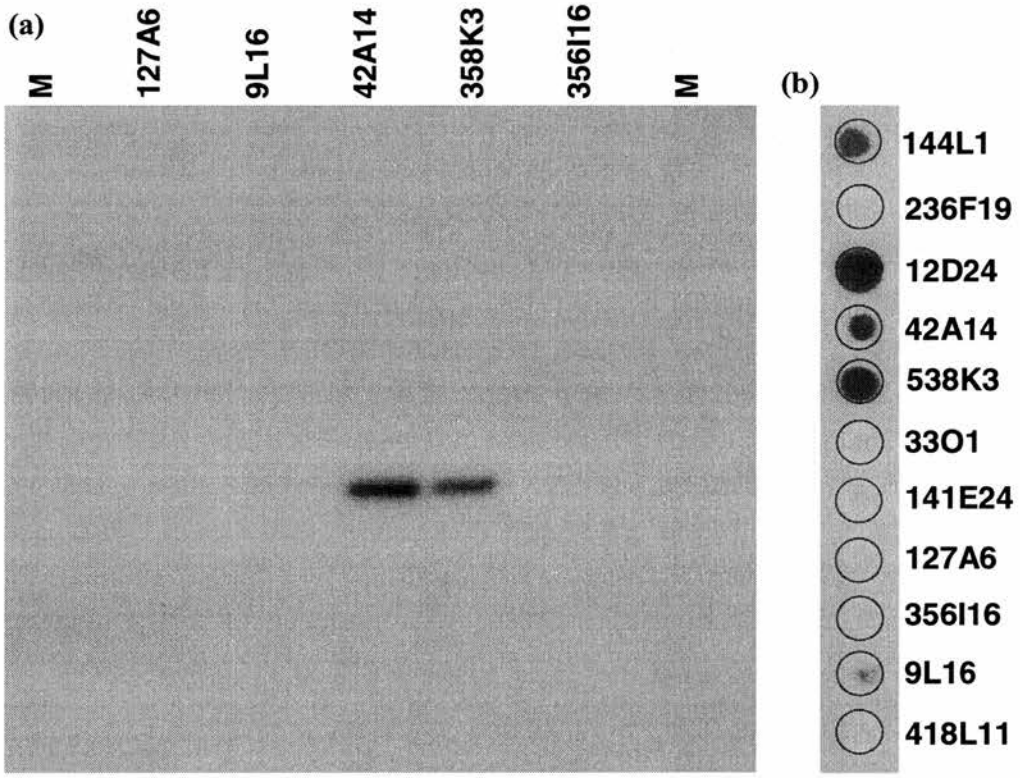


Figure 4.7; Mouse BAC clone validation. **(a)** Oligonucleotide hybridisation of mouse *DISC1* exon 13 sequence to a Southern blot of the *HindIII* restriction digest shown in figure 4.5 (panel 'b'). Clones 42A14 and 538K3 were both found to contain exon 13. Lanes marked 'M' indicate the position of marker lanes on the gel. **(b)** Spot blots of mouse BAC DNA hybridised with an oligonucleotide corresponding to mouse *DISC1* exon 9. Based on these results, clones 144L1, 42A14 and 538K3 were considered positive for exon 9 of *DISC1*.

Clone 12D24 gave a strong hybridisation signal, however, its restriction pattern was not consistent with the expected FPC pattern. The faint signal for clone 9L16 was considered spurious based on the relatively weak nature of the signal.

4.5 Identification of chicken BAC clones

Comparative sequence analysis benefits from the alignment of multiple homologous sequences, ideally separated by a range of evolutionary distances (section 1.7.1). As a resource for the future multiple sequence comparative analysis of the chromosome 1 breakpoint locus, chicken (*Gallus gallus*) BAC clones were isolated. Using the strategy that worked successfully for the isolation of the *Fugu* locus, a *TRAX* derived probe (insert of the d10-m12 cDNA clone, section 8.4.2) was used to screen a chicken whole genome BAC library (section 2.9.7). The d10-m12 cDNA clone insert was hybridised at high stringency (section 2.10.2) to gridded library filters, identifying 13 strongly hybridising clones. The identified clones were obtained and validated by restriction digestion, Southern blotting and hybridisation with the d10-m12 probe (figure 4.8). All 13 clones were found to hybridise the probe consistent with them containing a chicken homologue of *TRAX*. Further validation (sample sequencing) of these clones would be required to determine if they also contain a *DISC1* homologue.

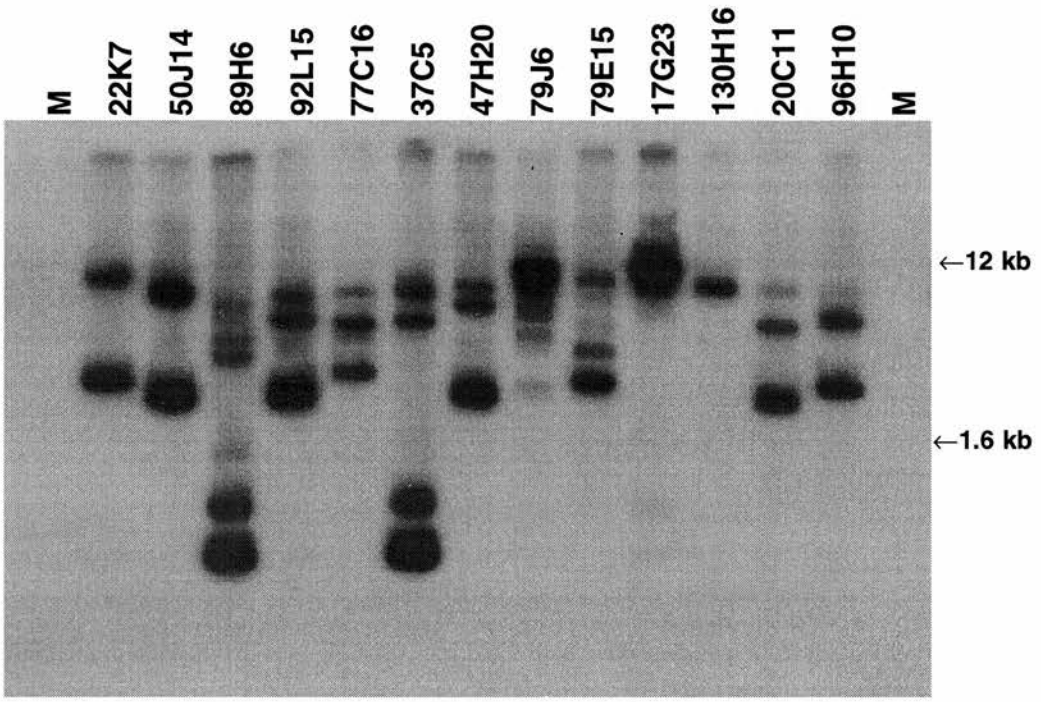


Figure 4.8; Validation of chicken BAC clones. Identified chicken clones were restriction digested (*EcoRI*), the fragments electrophoretically separated and Southern blotted. The Southern blot was hybridised at high stringency (section 2.10.2) with a double stranded DNA probe derived from the d10-m12 TRAX cDNA clone insert (section 8.4.2).

4.6 Discussion

A method for the integration of complementary human genome assemblies was developed and applied to the chromosome 1 breakpoint region to construct two blocks of contiguous sequence separated by a small gap of approximately known size. This assembled sequence was validated by internal controls and was consistent with gene and marker orders that had been experimentally determined. The sequence assembly forms the reference sequence for comparative genomic annotation and alignment developed in subsequent chapters. The method used for assembly was not fully automated and relied heavily in the later stages on manual intervention to direct the assembly. The logic of the manual intervention could be rigorously defined and incorporated into an automated method. However, limits on the use of some data sets

and the changing nature of draft human sequence suggest that the benefits of such an automated system would be very short lived.

Using the assembled human genomic sequence as a primary resource, a mouse BAC contig was constructed across the chromosome 1 breakpoint orthologous region from mouse. This strategy was based on the principal of identifying reciprocal best matches (RBMs) between the assembled human sequence and mouse BAC end sequences. These RBMs were used to anchor the FPC contigs to the human assembly allowing overlapping FPC contigs to be merged with confidence and a relative orientation of the contigs to be determined. This methodology was found to be rapid, efficient and not readily undermined by highly similar sequences such as processed pseudogenes. This approach could be scaled up to whole genome scale comparisons where there is a good reference (assembled) genome for the comparison. This system not only provides information about the clones over a particular orthologous region of interest but can also indicate boundaries of conserved synteny and can be used to estimate the relative size of orthologous genomic regions. It should be noted that the performance of this method has not been tested on regions of the genome containing closely related multi-gene families such as the odorant receptor or zinc finger gene clusters.

From the mouse FPC contig across the region and anchors in the form of RBMs between the human sequence and the FPC contig, a set of clones was identified that was likely to contain the entire *TRAX* and *DISC1* genomic region. These clones were validated experimentally and submitted for sequencing. The sequence of these clones will be of use for further comparative genomic analysis of the chromosome 1 breakpoint region and provide the basic information necessary to construct mouse models of the translocation region (discussed in chapter 10).

The genes in the general region of the chromosome 1 breakpoint were identified as were the genes in the orthologous region of the *Fugu* genome defining a boundary of conserved synteny. The observation that the *TM7SF1* and *NID* genes identified in the *Fugu* clone contig are also located on chromosome 1 in the human suggest that there

may be an extended region of conserved synteny with localised disruption of gene order and orientation. Such localised disruptions of gene order and orientation have been previously reported in human versus *Fugu* genomic sequence comparisons (Armes *et al.,* 1997) and have been proposed as a frequent event in the evolution of all genomes (Seoighe *et al.,* 2000).

Based on the proximity to the chromosome 1 breakpoint and the availability of orthologous *Fugu* genomic sequence for cross species genomic comparison a region of primary interest was defined. The region of primary interest contained the identified genes *EGLN1*, *TRAX* and *DISC1* was well as the *DISC2* transcript. Each of these genes and the flanking genomic sequence is investigated in detail in the subsequent chapters. Other genes identified in the wider genomic context of the chromosome 1 breakpoint were evaluated as functional candidates for a significant role in the psychiatric phenotype of translocation carriers. The *GNPAT* and *EXO84* genes are both likely to be essential for survival, although mutations in the *GNPAT* gene causing a reduction of the enzyme activity of the associated gene product are associated with a known human syndrome (section 4.2.3). Neither the *GNPAT* nor *EXO84* are considered good candidates for a role in the susceptibility to major mental illness.

The *KIAA1389* and *Q9BSD7* genes both possess known protein domains that allowed at least some aspects of their functional role to be predicted. KIAA1389 was predicted to be a member of the Rap_GAP family of proteins involved in intracellular signalling. The Q9BSD7 predicted protein was found to be almost entirely composed of an AAA nucleic acid binding domain. The AAA domain is an ancient and highly conserved domain that has been acquired for several roles through the course of evolution although it is often associated with a chaperone like activity. For example, AAA domains are found in the ABC transporter family of proteins including CFTR where they are involved in the opening and closing of transmembrane channels through the induction of protein conformational changes. As for their functional candidacy for a role in the t(1;11) phenotype, the *KIAA1389*

and *Q9BSD7* genes cannot be considered outstanding candidates without further functional information regarding the pathways in which they are involved.

For the genes *DJ876B10.3* and *KIAA1383* there is no functional information available, consequently these genes cannot be evaluated at this time in terms of functional candidacy.

The *PCNXL2* gene encoding a homologue of the *Drosophila* neurodevelopmental gene *Pecanex* can be considered an excellent functional candidate for a role in the mental illness and P300 ERP (section 1.6) phenotypes of the t(1;11) carriers. Within *Drosophila*, *Pecanex* has a strictly neurological expression pattern and null mutants have severe disruption of neural development (LaBonne *et al.,* 1989). The observed similarity with *Notch* mutant embryos (LaBonne *et al.,* 1989) is of particularly significant interest as *NOTCH4*, one of the four human homologues of the Drosophila *Notch* gene has been significantly ($P < 0.000036$) associated with a susceptibility to major mental illness with a core phenotype including schizophrenia (Wei and Hemmings, 2000). However, four follow up studies of these finding for *NOTCH4* association have failed to find supporting evidence for this locus in other population samples (Ujike *et al.,* 2001; McGinnis *et al.,* 2001; Sklar *et al.,* 2001 Imai *et al.,* 2001). The functional candidacy of *PCNXL2* is discussed further in chapter 10.

Chapter 5

Genomic sequence annotation

5.1 Preface

A total of 0.8 Mb of human sequence was assembled over the chromosome 1 breakpoint region (section 4.3.2) and the previously known *DISC1*, *DISC2* and *TRAX* genes were identified. A novel homologue of the *C. elegans* gene *Egl-9* was also identified (section 4.2.1). This small fraction (0.026%) of the human genome is still larger than many bacterial genomes and could potentially encode many more than the three protein coding genes identified using the crude homology based method described in section 4.2.1. As well as the possibility of novel protein coding genes, non-coding RNAs and previously unidentified exons of the known genes may remain undetected in the genomic sequence. The orthologous region from *Fugu* (section 3) represents an important resource for identifying these features. To complement the *Fugu* versus human comparative genomics approach, the preliminary computational annotation of both the human and *Fugu* genomic sequence was undertaken. The strategy used for preliminary annotation of genomic sequences are described and the findings presented. Resources for the annotation of human sequences are particularly plentiful in the form of over 3 million expressed sequence tags (ESTs) and *ab initio* gene finding software that was specifically trained on human genes (section 2.11.2). Databases of known human repetitive sequences (section 2.11.1) also represent an important resource for the annotation of human genomic sequence.

Whole genome shotgun sequence data from the mouse genome is also available. While this sequence is unassembled it does have the potential to highlight particularly conserved regions such as protein coding exons. Similarly there is

whole genome shotgun sequence available for the *Tetraodon* genome that has great potential for the annotation of *Fugu* sequence.

Comparative sequence annotation between *Fugu* and human chromosome breakpoint regions was a key aim for the work undertaken in this thesis (section 1.8). Prior to the availability of the *Fugu* or contiguous human sequence over the region, the tools and methodologies for genomic sequence comparison between these distantly related vertebrates were optimised using a model locus. This model locus was that of the cystic fibrosis transmembrane-conductance regulator (*CFTR*). The biological findings from this comparative analysis of the *CFTR* locus are presented in brief (see Davidson *et al.,* 2000 for complete results). The methodological approaches used during comparative analysis of the *CFTR* locus were further developed into a novel approach for the alignment of evolutionarily divergent genomic sequences.

5.2 Preliminary annotation of human genomic sequence

The strategy (pipeline) used for preliminary human genomic annotation is outlined in figure 5.1. The annotation pipeline identified three processed pseudogenes, two spliced transcripts of undetermined protein coding status and multiple EST clusters in the introns of *DISC1*. Regions of apparently non-coding conservation were also identified in mouse whole genome shotgun sequence data.

5.2.1 Processed pseudogenes

In addition to the previously identified genes: *EGLN1*, *TRAX*, *DISC1* and *DISC2,* sequences similar to the known proteins, (a) "small nuclear ribonucleoprotein D2" (SPTR: AAH00486) and (b) APE0615 (SPTR: Q9YEG1) were identified. Both of these sequences exhibiting sequence similarity in translation to known proteins were located in the intergenic space between *TRAX* and *EGLN1*. Neither of these sequences maintained an open reading frame, and both exhibited substantial nucleotide similarity with genes elsewhere in the genome (data not shown). These observations suggested that both represented processed pseudogenes. These probable pseudogenes were not represented by ESTs suggesting that they are not expressed sequences.

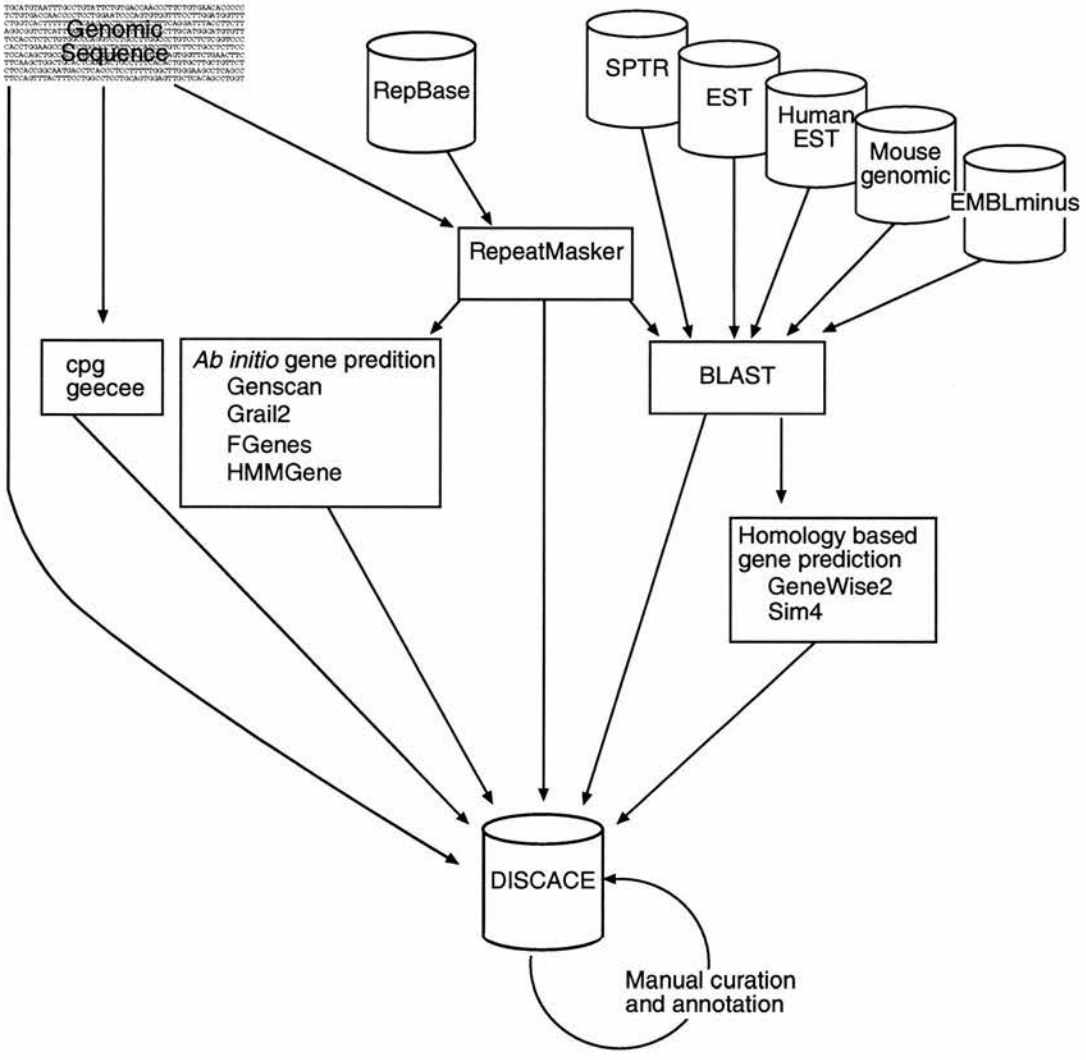


Figure 5.1; Pipeline for preliminary annotation of human genomic sequence. Databases are indicated by cylinders. RepBase and SPTR are described in section 2.11.1. EMBLminus, EST and Human EST are sub-sets of the EMBL database described in section 2.11.1. The mouse genomic database refers to a novel BLAST database constructed from unassembled mouse whole genome shotgun sequence combined with mouse BAC-end sequence and mouse genomic sequence from the EMBL database. The software indicated is described in section 2.11.2. BLAST was used in BLASTN mode for nucleotide databases and BLASTX mode for protein databases. The mouse genomic database was searched using both BLASTN and TBLASTX modes. GeneWise2 was used with the modifications described in section 2.11.2. DISCACE is an AceDB database (section 2.11.2) developed in part from data models provided by S. Morris (Edinburgh University). The manual curation and annotation refers to the investigations and findings described throughout this thesis. Figure 5.2 provides an example of data populating the database.

5.2.2 Expressed sequences

A total of 1,198 EST sequences showed significant (BLASTN E-value <0.0001) sequence similarity to the repeat masked human genomic sequence from the region of interest. Many of these were single unspliced ESTs that are likely to represent genomic contamination of cDNA libraries and unprocessed heterogeneous nuclear RNA. Also, ESTs were found to align with less than 98% identity in non-repetitive sequence to processed pseudogenes within the genomic region (section 5.2) suggesting that the ESTs did not represent expression of the pseudogene but rather the real gene. Consequently, human ESTs were not further evaluated unless they were spliced or represented by multiple overlapping ESTs. A further requirement was that the EST aligned to genomic sequence with ≥98% identity for more than 40 consecutive nucleotides of non-repetitive sequence (minimum E-value of $<2\times10^{-12}$ when using the human EST database). Multiple ESTs representing transcripts of the human M-phase phosphoprotein homologue (EMBL: AF100742) aligned to genomic sequence between *TRAX* and *DISC1*. However, the identity of the alignment was 96% with multiple small deletions. Only the 3' end of the 3' UTR aligned and there was no sequence similarity with the protein coding region of the transcript AF100742. This sequence is therefore likely to represent a retrotransposition of the 3' UTR of AF100742.

EST clusters that were found to lie within the introns of *DISC1* were investigated experimentally and are considered in chapter 6. Upstream of *DISC1*, two EST clusters were identified, both being represented by multiple spliced ESTs that were >98% identical in alignment to the assembled genomic sequence. These EST clusters are subsequently referred to as *Backtrax* and *Foretrax* for reasons that will be described.

Backtrax

Searching of the assembled human genomic sequence (section 4.3.2) against EST databases identified multiple ESTs representing transcription from the *TRAX* locus that appeared to extend the 5' UTR of *TRAX*. However, several of the cDNA clones from which the ESTs were derived had been directionally cloned allowing the

direction of transcription to be inferred. ESTs that could be oriented, suggested that they represented transcription from the opposite strand to *TRAX* encoding transcripts. This conclusion was supported by alignment of EST sequences to the assembled genomic sequence. Sequence flanking the 20 splice sites (multiple alternative splicing events) did not conform to known splice consensus sequences if the *TRAX* orientation of transcription was assumed. However, if the opposite transcriptional orientation was assumed, every splice site was a good match to the AG-GT splice consensus sequences (Moore, 2000). Therefore, these ESTs represent a transcript originating at the *TRAX* promoter, transcribing away from the *TRAX* gene (figure 5.2).

Iterative assembly of these ESTs (section 2.11.4) has resulted in the definition of a transcript with at least four different splice variants (figure 5.2). These transcripts have been tentatively named "*Backtrax*" in reference to their inverse transcriptional relationship to the *TRAX* gene. All four of the *Backtrax* transcripts utilise consensus splice sites and poly adenylation signals. It was interesting to note that the transcripts spliced into and out of interspersed repetitive elements and that only the first exon (one transcript) or first and second exons (three transcripts) of the *Backtrax* transcripts represented non-repetitive sequence (figure 5.2).

The largest open reading frame for each transcript ranges from 57 codons to 75 codons. The candidate translation initiation site for the longest open reading frame of each transcript was in an adequate context for translation initiation (Kozak, 1996). The potential open reading frames had no detectable similarity to known proteins by BLASTP searching (section 2.11.2) or through screening against the PFAM protein domain database (section 2.11.1). The protein coding status of these transcripts is unclear, although splicing after the stop codon and use of interspersed repetitive elements as exons suggest that these are probably not protein coding transcripts. The protein coding status is considered further in comparison with *Fugu* genomic sequence (section 8.5.7).

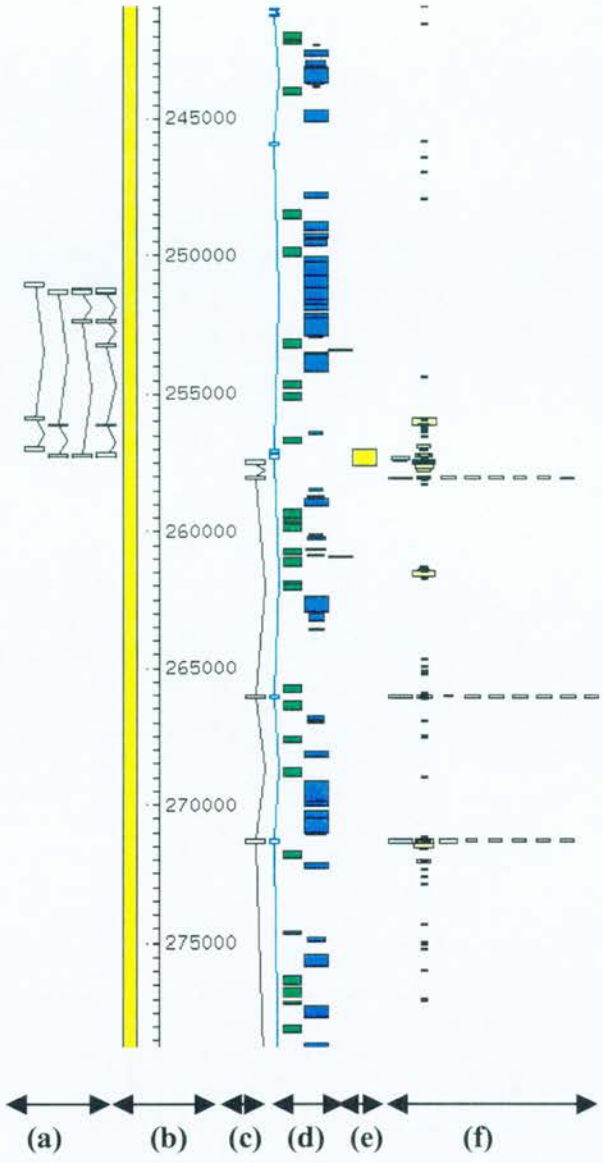


Figure 5.2; Screen shot from the DISCACE AceDB database (section 2.11.#) showing the 5' end of *TRAX* and the Backtrax transcripts. **(a)** Alignments of the assembled Backtrax transcripts to genomic sequence. As shown, Backtrax is transcribed from the bottom to the top. Boxes indicate exons, angled lines introns. **(b)** The yellow bar indicates contiguous genomic sequence, and adjacent numbers the nucleotide coordinates of that sequence. **(c)** Alignment of the human *TRAX* transcript to genomic sequence. Exons 1 to 4 are shown. **(d)** The blue boxes joined by thin lines indicate a "genscan" (section 2.11.2) prediction of a gene that accurately predicts exons 3 and 4 of *TRAX*. Green and blue solid boxes indicate short and long interspersed repetitive elements respectively, detected by RepeatMasker (section 2.11.2). **(e)** The yellow box indicates a CpG island prediction encompassing the first exons of Backtrax and *TRAX*. CpG island predictions were carried out using the cpg

program (section 2.11.2). **(f)** Boxes indicate homology to known sequences, the width of the box corresponds to the BLAST bit score of the alignment. Light blue boxes indicate TBLASTX (section 2.11.2) detected sequence similarity to proteins in the SPTR database (section 2.11.1). Orange boxes indicate sequence similarity to human ESTs as detected by BLASTN (section 2.11.2). Light pink boxes on the right of the figure indicate sequence similarity detected between the human genomic sequence and mouse whole genome shotgun sequences (section 2.11.1).

Foretrax

The eleven ESTs represented by the UniGene cluster Hs.25555 (section 2.11.1) align with 98 to 100% identitity to the genomic sequence between *TRAX* and *DISC1*. Analogously to the naming of the *Backtrax* transcript, this transcript is subsequently referred to as "*Foretrax*". Assembly of these ESTs (section 2.11.4) suggests that they represent a single transcript of 612 nucleotides that is comprised of two exons and utilises a single consensus polyadenylation signal. Nine of the eleven ESTs were derived from the same mammary gland cDNA library. In total this transcript was only represented by five independent cDNA clones. The three cDNA clones for which there is 5' end sequence available (R71937, R48416, AI820788) all aligned to the same nucleotide coordinate in genomic DNA, indicating that this represents the transcript start site for the *Foretrax* transcript. The longest open reading frames of this transcript are 84 and 79 codons in length. The hypothetical start codons for both open reading frames are in an adequate context for translation initiation (Kozak, 1996) although both have shorter upstream open reading frames. There was no identifiable homology to known proteins or protein domains. The protein coding status of the *Foretrax* transcript is considered further in comparison with *Fugu* genomic sequence (section 8.5.7).

TRAX – DISC1 intergenic splicing

While performing 5' rapid amplification of cDNA ends (RACE) to define the 5' end of the *DISC1* transcript, Millar *et al.,* (2000) identified intergenic transcripts between the *TRAX* and *DISC1* gene. Subsequent investigation demonstrated that intergenic splicing could be detected by RT-PCR between exon 5 and exon 6 of TRAX and exon 2 of *DISC1* with the variable inclusion of three intergenic exons (Millar *et al.,* 2000b). An open reading frame was not conserved between *TRAX* and

DISC1. Intergenic splicing was also detected by RT-PCR between *TRAX* and *DISC1* in the mouse although intergenic exons were not found in these transcripts (R. Devon, personal communication). Again, the intergenic transcripts would not maintain a continuous open reading frame between the *TRAX* and *DISC1* open reading frames. The intergenic splicing of *TRAX* and *DISC1* is considered further in comparison with *Fugu* genomic sequence in chapter 8.

5.2.3 Evolutionarily conserved sequences

BLASTN and TBLASTX (section 2.11.2) searching of the human sequence contigs against the mouse whole genome shotgun sequence identified multiple regions of sequence similarity. The E-value of 0.001 was used as an arbitrary and liberal threshold for incorporation into the DISCACE database for subsequent evaluation. The sequence similarity observed did not indicate any additional protein coding exons. Many of the previously defined exons of *EGLN1*, *TRAX* and *DISC1* were found to be represented in the mouse sequence. Mouse sequences representing a homologue of *DISC1* were identified and localised assemblies produced to determine the full open reading frame of mouse *DISC1* (section 6.4.3). Of particular note was a single mouse sequence (G10P6460721RA4) that did not overlap other sequences in the mouse database. This sequence was found to align with 93% identity (compared to an average 76% nucleotide identity between mouse and human *DISC1* coding sequences) to human sequence from *DISC1* intron 10 (coordinates 368569 to 369141 of the human assembly, appendix I). This sequence is not repetitive and has no detectable similarity to known interspersed repetitive elements. Relative insertions and deletions and the absence of a substantial uninterrupted reading frame in either human or mouse sequence indicates that this does not represent protein coding sequence. Non-coding conservation in the *DISC1* region is investigated further in section 6.6.1.

5.3 Preliminary annotation of *Fugu* genomic sequence

The preliminary annotation of *Fugu* genomic sequence was carried out in a similar manner to the annotation of human genomic sequence (summarised in figure 5.1). The principal differences between approaches used for human and *Fugu* sequence annotation reflected the high repeat content of the human genome and low level of

interspersed repeats in the *Fugu* genome. *Ab initio* gene prediction and homology to known proteins indicated that the *Fugu* sequence contig did not encode any protein coding genes other than the *EGLN1*, *TRAX* and *DISC1* genes that had previously been identified (section 4.2).

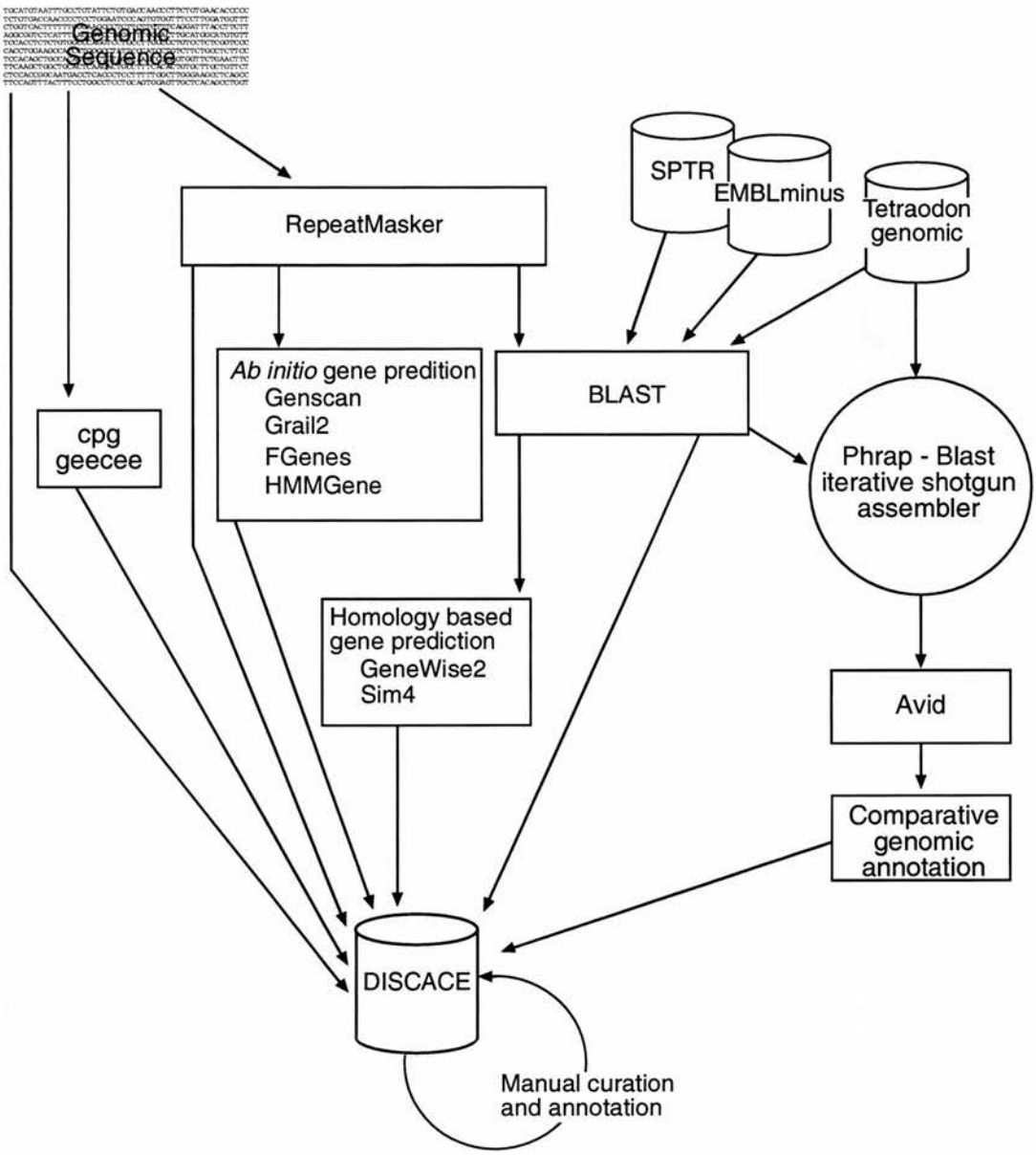


Figure 5.3; Pipeline for preliminary annotation of *Fugu* genomic sequence. Databases are indicated by cylinders. RepeatMasker was used to mask only low complexity sequence, a database of interspersed repetitive elements was not used. The Phrap – Blast iterative shotgun assembler is described in section 2.11.4. Avid was used to filter, order and orient sequence contigs produced by the assembler. The software indicated is described in section 2.11.2 and databases in section 2.11.1. BLAST was used in BLASTN mode for nucleotide databases and BLASTX mode for protein databases. DISCACE is an AceDB database developed in part from data models provided by S. Morris (Edinburgh University). The manual curation and annotation refers to the investigations and findings described throughout this thesis.

5.3.1 Repeat sequences

Low complexity and interspersed repetitive elements as well as a surprising number of novel tandem repeat arrays were identified in the *Fugu* genomic sequence contig. Within the *Fugu* sequence contig, there were 39 low complexity regions (RepeatMasker, section 2.11.1) with a mean length of 100 nucleotides and a median of 78 nucleotides. Tandem repeat arrays with a basic repeat unit length of between 5 and 50 nucleotides and between 2 and 157 copies were found to be abundant in the *Fugu* genomic sequence contig (figure 5.4). It was also noted that the number of repeat copies differed in regions of overlap between the two cosmid clones sequenced (section 3.7). These discrepancies indicate that the repeats are polymorphic in the *Fugu* population, although the possibility of changes in repeat length during cosmid clone propagation in *E. coli* cannot be ruled out. Table 5.1 summarises the tandem repeats identified (etandem, a component of the EMBOSS package, section 2.11.2).

Previous descriptions of *Fugu* genomic sequences have typically not found such an abundance of tandem repeat sequences (Baxendale *et al.,* 1995; Aparicio *et al.,* 1997; Armes *et al.,* 1997; Miles *et al.,* 1998; Davidson *et al.,* 2000). The orthologous human sequence is devoid of such large repeat arrays, there are only three tandem repeat arrays with more than two tandem repeat units in the sequence contig containing *DISC1*, the largest of those is four tandem copies of 30 nucleotide basic repeat unit. Assembled *Tetraodon* sequence over the *TRAX – DISC1* region (section 5.3.2) is completely devoid of such tandem repeat arrays (figure 5.3 and data not shown). However, gaps in the *Tetraodon* sequence contig correspond in many cases to the location of tandem repeat arrays in *Fugu* genomic sequence (figure 5.5) indicating that there may be a gross under representation of these sequences in the *Tetraodon* whole genome shotgun sequence. The initial shotgun sequencing of *Fugu* cosmid clones showed substantial under representation of these repeat arrays. A systematic under representation of a subset of sequences was the principal argument against application of the whole genome shotgun sequencing strategy to higher eukaryotic genomes (Green, 1997).

Searching of the *Fugu* sequence against the EMBL database (section 2.11.1) suggested there were two interspersed repetitive elements within the sequence (data not shown). These repeats are subsequently referred to as Rep1 and Rep2 (coordinates of 8954 to 9112 and 37434 to 37500 of the *Fugu* sequence contig respectively). Searching (BLASTN, section 2.11.2) against *Fugu* and *Tetraodon* whole genome shotgun sequence indicates that there are at least 250 copies of the Rep1 repeat in the *Fugu* genome and there is a related repeat in the *Tetraodon* genome (data not shown). Preliminary analysis suggests that the 67 nucleotide sequence referred to as Rep2 represents a fragment of a larger interspersed repetitive element. The Rep1 and Rep2 repeats do not show detectable sequence similarity with each other, or previously described repeat elements. The Rep1 and Rep2 repeats were masked prior to the iterative assembly of *Tetraodon* genomic sequence (section 5.3.2) but were unmasked for comparative analysis with non-puffer fish sequences.

Start[a]	End[b]	Unit length[c]	#[d]	%ID[e]	Repeat unit sequence[f]
1	378	21	18	92.6	tataggaggtgggacggggtc
2234	2299	11	6	77.3	cgcggcggcct
2352	2492	47	3	94.3	gtcgcggcggcctcatgttggcggcctcgtggcggcggcggcctcat
5874	5927	18	3	98.1	gtcacccgcgccaggtgt
10714	11289	48	12	98.6	ctcggctgggtgcacggagctgtgggtaggggggcgctctccacgtgac
14276	14667	49	8	75.0	gggttaggttagggttaggttagggtagggttaggttaggttagggtta
14704	14891	47	4	84.6	gttatggttaggttaggctaggttaggttggggttaggttaggctag
15306	15365	15	4	100.0	gccaacatgggggc
19816	19905	45	2	90.0	atcatcccggcccagatcatcccggcccagatcatctcggcccag
19913	20092	15	12	91.1	gcccagatcatctca
21006	21085	20	4	100.0	gggggccgtgggggccaggc
21593	21760	21	8	63.1	gagcagctgaaggagcagctg
22253	22336	12	7	88.1	gtgcgtgcgtgc
22751	22870	30	4	98.3	agtgcacccgggtcgtcactagggggcgct
25453	25507	11	5	89.1	ctgccccccctc
27293	27682	15	26	93.3	tgggagagatcttca
27987	28046	12	5	90.0	tcattccttcat
29097	29300	12	17	91.2	agggttagggtt
29465	29600	17	8	74.3	cctgtgggctactaacc
30661	31050	39	10	72.3	tgtaccccggggctctactggctcccccaggtcgtgtct
32580	32931	16	22	77.0	ctcctctcctcctctc
33210	33269	12	5	95.0	ggccccccgccgt
34215	34277	21	3	90.5	cccagggggccttaggggggc
34640	34889	50	5	99.6	tcacctggttagtctcaggaccaggagatgacctgaacgtccataacggg
35179	35334	12	13	94.9	tgtctgtctgtc
35350	35545	49	4	90.3	tcacctgtagatgttcctctcacctgtgcagttgtgcaggttccaccgg
36048	36195	37	4	83.8	gggggggggtaacgggggggggggcgtaacgggaggg
36416	36714	23	13	95.7	ggagcatcttggttctggacccg
36877	37437	11	51	92.7	gttagcaccag
37612	38227	11	56	91.2	cagggtagcac
38274	38828	37	15	60.0	gggggaaggggggggaggggagaagaggggaggggg
38939	39033	19	5	72.6	cactcccccactcccccctg
41391	41618	12	19	95.2	gcaacagacagg
41975	42018	11	4	97.7	cctggaaatca
43468	43597	10	13	93.1	acacacacac
44617	44814	18	11	98.5	gctccagcgtggtgagga

Table 5.1; Tandem repeat arrays in the *Fugu* sequence contig. Tandem repeat arrays detected by etandem (section 2.11.2). **(a)** Start coordinate of tandem repeat in *Fugu* sequence contig. **(b)** End coordinate of tandem repeat in *Fugu* sequence contig. **(c)** Length in nucleotides of the basic repeat unit. Several of the repeats units have an internal repetitive structure. **(d)** The number of tandem copies of the basic repeat unit within the repeat array. **(e)** The average percent identity of each repeat in the array from the repeat unit consensus sequence. **(f)** Nucleotide sequence of the basic repeat unit.

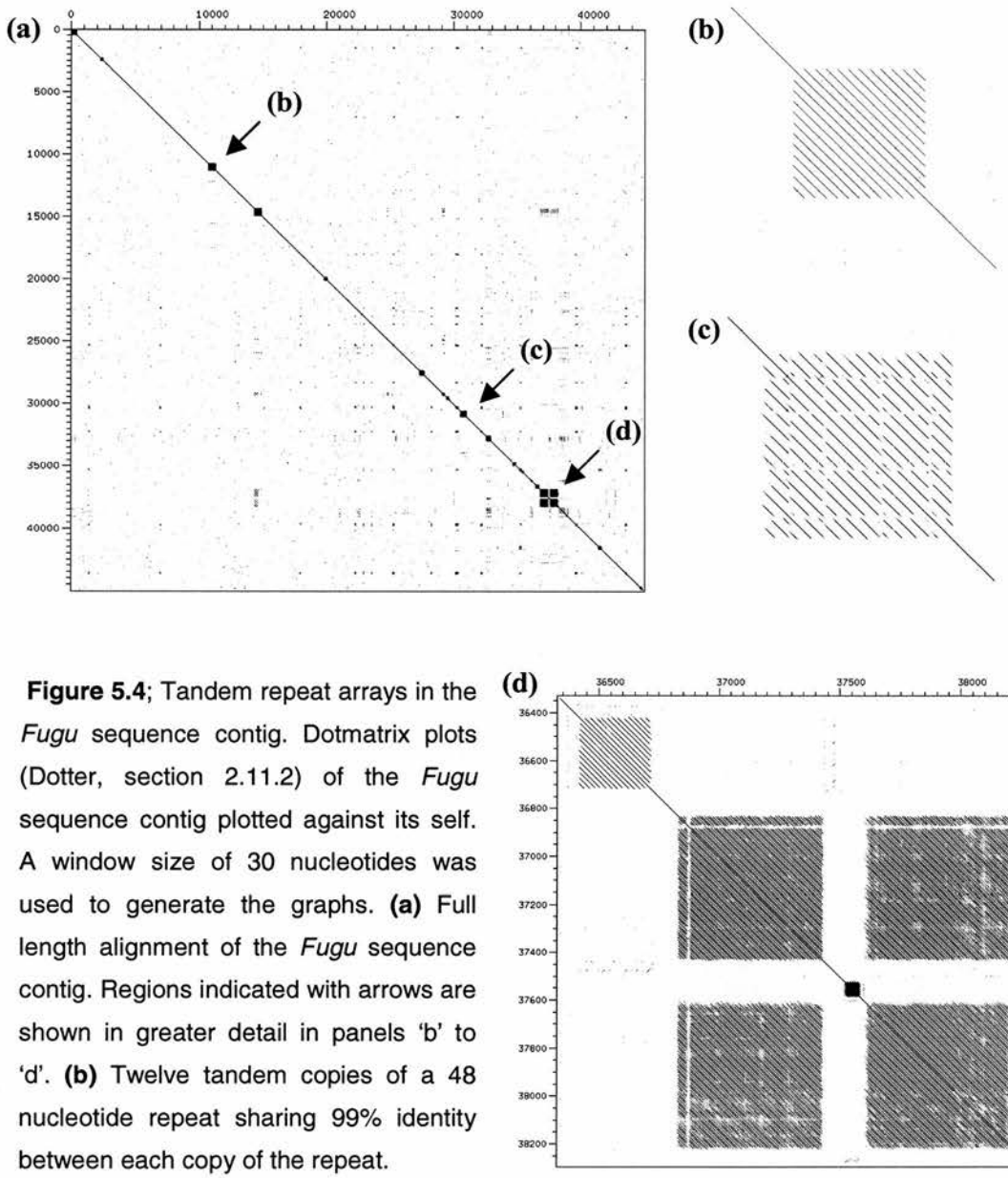


Figure 5.4; Tandem repeat arrays in the *Fugu* sequence contig. Dotmatrix plots (Dotter, section 2.11.2) of the *Fugu* sequence contig plotted against its self. A window size of 30 nucleotides was used to generate the graphs. **(a)** Full length alignment of the *Fugu* sequence contig. Regions indicated with arrows are shown in greater detail in panels 'b' to 'd'. **(b)** Twelve tandem copies of a 48 nucleotide repeat sharing 99% identity between each copy of the repeat. **(c)** Ten tandem copies of a 39 nucleotide repeat (72% identity between repeats) with an apparent higher order structure indicating multiple rounds of duplication and diversification. **(d)** A complex of repeats that caused particular problems in sequence assembly (section 3.5.6). The upper left repeat is thirteen tandem copies of a 23 nucleotide repeat. The large interrupted block of repeat is two arrays of an 11 nucleotide repeat with 56 copies in the block shown on the left and 51 copies in the block on the right. There is 92% identity between repeats. The 11 nucleotide repeat units are themselves organised into a higher order repeat of 4 consecutive repeats identified by a repeating pattern of single nucleotide substitutions. The large repeat array is interrupted by the insertion of an interspersed

repetitive element (Rep2) which itself is interrupted by the expansion of a low complexity sequence.

5.3.2 Annotating *Fugu* with *Tetraodon*

An additional resource available for the annotation of *Fugu* genomic sequence was the approximately 2.5× whole genome shotgun sequence from the puffer fish *Tetraodon nigroviridis* (section 2.11.1). BLASTN searching of the *Fugu* genomic sequence contig against the *Tetraodon* whole genome shotgun sequence suggested that there was potential to perform a localised sequence assembly. The contiguous *Fugu* genomic sequence (section 3.5.6) was used to initiate the interactive assembly process described in section 2.11.4. Similarity to the contiguous *Fugu* sequence was also used to order and orientate the subsequently generated sequence contigs. In the final assembly, 70% of the *Fugu* genomic sequence aligned with the assembled *Tetraodon* sequence. (figure 5.5). The assembled *Tetraodon* sequence was utilised in the comparative analysis of individual gene regions described in subsequent chapters.

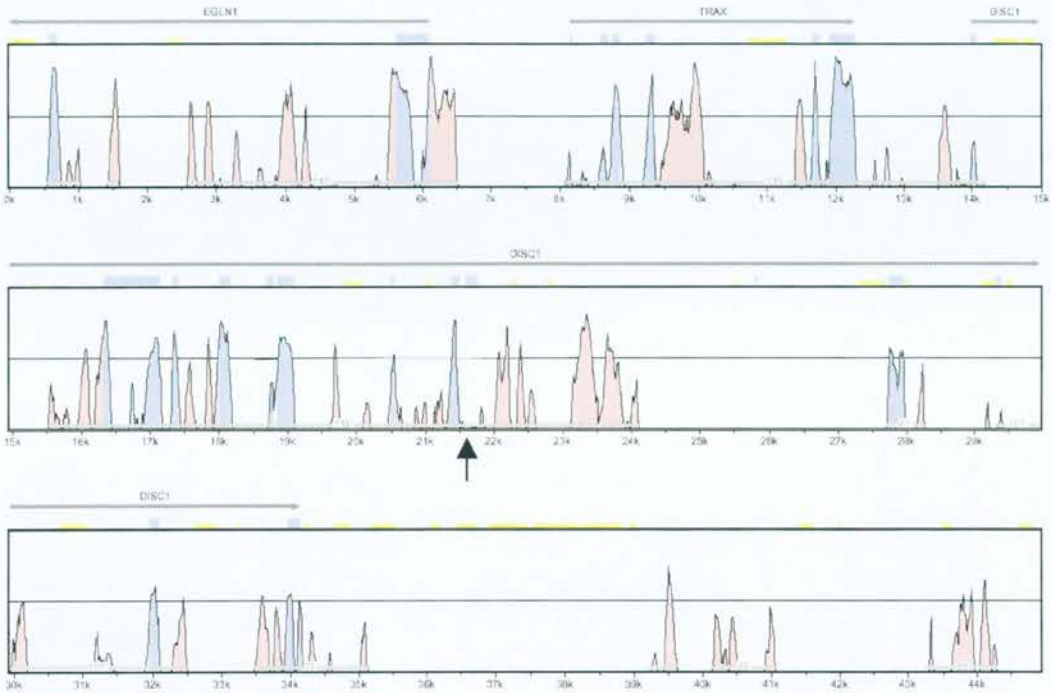


Figure 5.5 Alignment of contiguous *Fugu* genomic sequence to assembled *Tetraodon* sequence. Vista plot of Avid alignment (section 2.11.2) between assembled *Tetraodon* sequence and the *Fugu* sequence contig. The vertical axis represents 50% to 100% sequence identity in alignment, averaged over a sliding window of 100 nucleotides The horizontal midline through the graph indicates 75% sequence identity. Protein coding exons are annotated as blue boxes, the corresponding region of the graph is also coloured blue. Non-coding conserved sequences are indicated in pink on the graph. The annotation is derived from the results presented in chapters 6, 8 and 9. Yellow boxes above the graph indicate the extent of tandem repeats summarised in table 5.1. Horizontal grey bars indicate the extent of individual *Tetraodon* sequence contigs assembled from whole genome shotgun sequence (section 2.11.4). A black arrow indicates the position of DISC1 exon 9 that is referred to in section 5.5.

5.4 Human : *Fugu* genomic sequence alignment

Prior to the isolation and sequencing of the *Fugu TRAX – DISC1* locus (chapter 3), methods of comparative genomic alignment and analysis were developed using the *Fugu* and human *CFTR* locus as a model. The medical relevance of this locus is briefly put in context and the key findings summarised.

5.4.1 Introduction to the *CFTR* locus

Cystic fibrosis (OMIM: 219700) is a recessive monogenic disease of humans caused by the deficiency of a transmembrane chloride ion channel, the CFTR protein. The disease is characterised by a range of symptoms including meconium ileus (blockage of the gut) liver disease, pancreatic insufficiency, male infertility (bilateral absence of the vas deferens) and pulmonary disease (OMIM: 219700 for review). It is the pulmonary disease that is of devastating consequence to the sufferers of cystic fibrosis their life expectancy reduced to little over 30 years with currently available medical treatment (Doull, 2001).

The cause of pulmonary disease in cystic fibrosis is not fully understood but clearly relates to chronic infection of the lungs by pathogenic bacteria, most notably *Pseudomonas aeruginosa* (Goldberg and Pier, 2000 for review). As there is no obvious developmental lung defect in cystic fibrosis, it has been proposed that if the lungs of patients could be augmented with functional CFTR protein, it could represent an effective therapy for this disease. Gene therapy approaches have been devised and undergone preliminary trials to achieve this goal (Caplen *et al.,* 1995; McLachlan *et al.,* 1996; Bellon *et al.,* 1997; Alton *et al.,* 1999), although as yet there have not been any therapeutic benefits from this treatment.

The *CFTR* gene is known to be expressed in a tightly regulated tissue and cell specific manner (Smith *et al.,* 1995). It is therefore likely that any successful gene therapy approach will need to recapitulate the expression of endogenous, wild type *CFTR* with the exogenously introduced gene therapy vectors. To this end, genomic context vectors are being developed that are intended to drive the expression of a

CFTR encoding transcript, making use of the regulatory elements of human *CFTR*. However, a prerequisite of this approach is that the *cis*-acting elements that regulate the transcription of *CFTR* are known. Some of these elements have previously been identified through DNaseI hypersensitivity and gel mobility shift assay (Smith *et al.*, 1995; Smith *et al.*, 1996; Nuthall *et al.*, 1999).

The genomic sequence including a *Fugu* orthologue of *CFTR* was isolated and sequenced (H. Davidson; Davidson *et al.*, 2000). Comparative alignment and annotation between the human and *Fugu CFTR* genomic sequences was undertaken with the aim of identifying conserved non-coding sequences and further characterising the CFTR locus.

5.4.2 Preliminary genomic annotation - gene finding

The *Fugu CFTR* gene structure was predicted using the killifish (*Fundulus heteroclitus*) CFTR amino acid sequence (SPTR: O73677) as a homology template for GeneWise2 (section 2.11.2). The predicted gene structure (figure 5.6) showed the same 26 exon structure as human and mouse *CFTR* with each of the introns in the same relative position and splicing phase as the mammalian orthologues. *Fugu* homologues of the known mammalian genes *CBP90* (cortactin binding protein 90), *WNT2* (wingless-type MMTV integration site family member 2) and the unnamed gene represented by the transcript EMBL: Z43555; were all identified by BLASTX homology searching against the SPTR database. Each of these genes were also predicted in the human genomic sequence flanking the *CFTR* gene. Additional sequence similarity to the ankyrin protein domain (PFAM: PF00023) was identified (BLASTX against the SPTR database, sections 2.11.2 and 2.11.1) both upstream and downstream of the *CFTR* gene.

Ab initio gene prediction, sequence similarity to known proteins and human to *Fugu* genomic sequence alignment were manually integrated to predict the gene structures of two novel protein coding genes (subsequently *ANK1* upstream of *CFTR* and *ANK2* downstream). From the data available, it was not possible to confidently predict boundaries between *CBP90*, Z43555 and the *ANK2* gene. It is possible that they are all components of the same gene (Davidson *et al.*, 2000). *ANK1* was

predicted as a well defined single gene composed of 11 protein coding exons. This predicted gene structure has subsequently been supported by cDNA clone sequences and EST sequences from human and mouse (data not shown).

5.4.3 Genomic sequence alignment

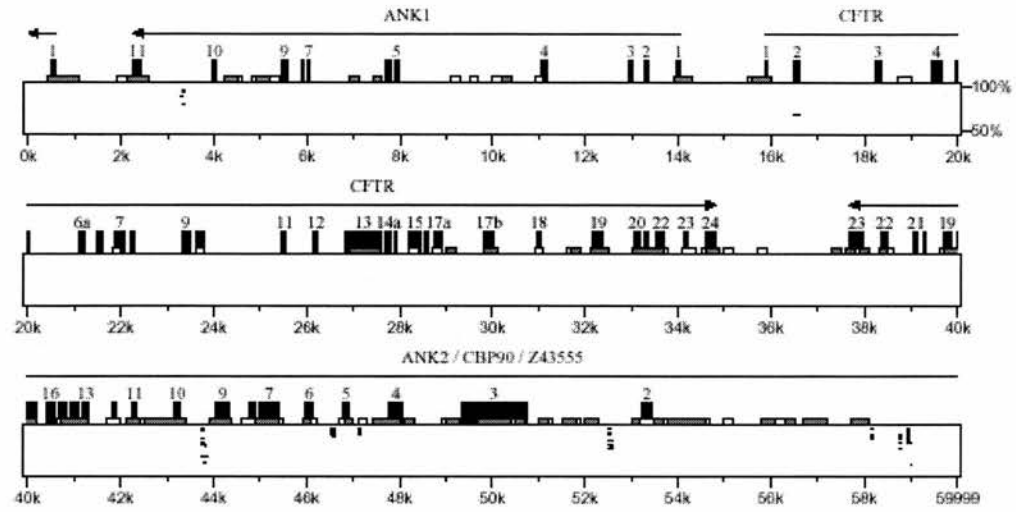
Initial genomic alignment between the human genomic sequence and *Fugu* genomic sequence (see Davidson *et al.,* 2000) was carried out using BLASTZ, a local alignment method (W. Miller, unpublished). Using default parameters, only one of the *CFTR* exons was identified even though there was known to be functionally relevant sequence conservation of these exons. The default parameters were extensively modified and tested (in consultation with Webb Miller, the author of the software). A set of parameters were defined that optimally detected the known functionally conserved sequences (*CFTR* exons). These parameters were: match = 1; transition = -0.6; transversion = -0.8 and alignment cut off = 18. However, no single set of parameters could be defined that would detect the homology for every exon (figure 5.6). Chaining and single coverage models (Schwartz *et al.,* 2000) were used to reduce background signals that largely reflected alignment between compositionally biased sequences.

The inability to detect known homology between *Fugu* and human genomic sequences suggested that the BLASTZ method may not have the sensitivity necessary to detect functionally conserved non-coding sequence and previously unknown exons. Other methodologies were investigated including rigorous Smith-Waterman alignment (Ssearch33, section 2.11.2), Needleman-Wunsch global alignment (needle component of EMBOSS, section 2.11.2), BLASTN (section 2.11.2) and ALIGN (section 2.11.2), none of which performed as well as BLASTZ with the optimised parameters (data not shown). TBLASTX (section 2.11.2) was found to be particularly effective at identifying conserved coding sequence although it was also extremely prone to finding spurious matches to low complexity sequence even when filtering of low complexity sequence was enabled. However, TBLASTX would not be expected to be efficient at identifying conserved non-coding sequences as it performs alignments between translations of nucleic acid sequence, a procedure that is not meaningful for non-coding sequences.

To detect non-coding conserved sequences pairwise dotmatrix alignments were carried out between orthologous sub-sequences of the human and *Fugu* genomic sequences. Sub-sequences were defined by the splice donor and acceptor sites of exons and start or stop codons for "external" exons. For example to align human and *Fugu CFTR* intron 1, genomic sequence representing the start codon to the splice donor site of exon 2 was excised from the genomic sequence (section 2.11.3) of both species and dotter alignment carried out between these two sequences. These two sub-sequences contain the coding exon sequence from exons 1 and 2 as well as the complete intron 1 sequence from both species. These dotter alignments were carried out systematically for every intron and intergenic space between identified orthologous exons. Evaluation of alignments was by visual inspection of the dotmatrix and manual inspection of candidate conserved sequences.

From the systematic dotmatrix comparisons, conserved non-coding sequences were identified in the sequence upstream of the *WNT2* gene, within intron 9 of *CFTR* and in intron 1 of *CFTR* (figure 5.7; Davidson *et al.,* 2000). The conserved sequence in intron 1 was of particular interest as it overlapped a previously defined DNaseI hypersensitive site in humans that correlated with transcription of the *CFTR* gene (Nuthall *et al.,* 1999). Sequence was not found to be identifiably conserved between human and *Fugu* at the sites of other reported human DNaseI hypersensitive sites (Davidson *et al.,* 2000).

(a)



(b)

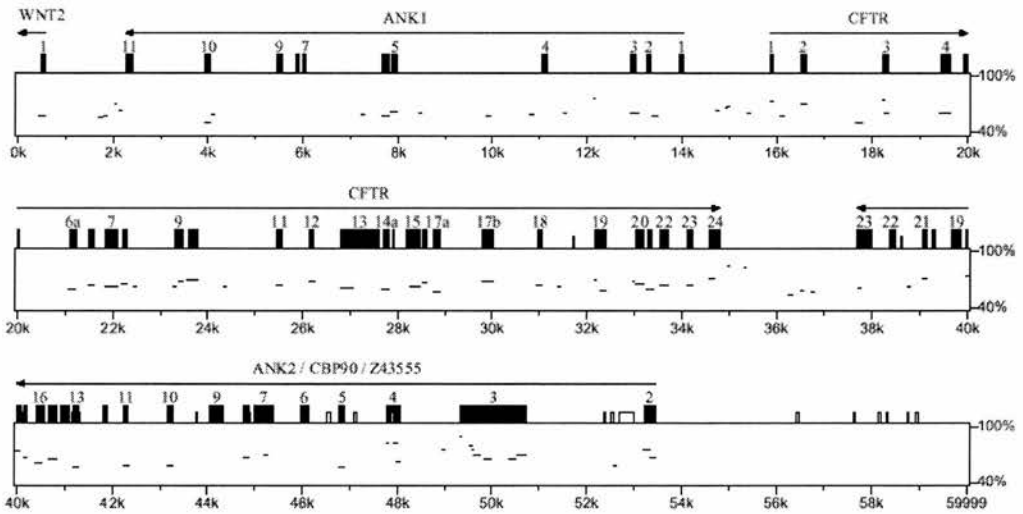


Figure 5.6; Human : *Fugu CFTR* percentage identity plots. Black boxes indicate predicted protein coding exons (section 5.4.2). The transcriptional orientation of genes is indicated by horizontal arrows. Exon numbers are marked where space permits. **(a)** Percentage identity plot generated using default PipMaker settings (section 2.11.2) **(b)** Optimised percentage identity plot for *Fugu* : human genomic sequence alignment. Percentage identity plot produced using the PipMaker software. With the exception of exon 6, nucleotide sequence similarity was found between all of the human and *Fugu CFTR* exons. Sequence similarity in genomic alignment was incorporated into gene predictions for the novel genes *ANK1* and *ANK2*. It was not possible to reliably predict if the *ANK2* gene represented one, two or there distinct genes (Davdison *et al.*, 2000).

5.4.4 A summary of findings at the *CFTR* locus

The *CFTR* promoter was found to be poorly conserved between species. Previously defined human promoter regulatory sequences such as Sp1 binding sites, Ap1 binding sites and negative transcriptional regulators were not found to be conserved in the *Fugu* sequence. The presence of a TATA box in the *Fugu CFTR* promoter and the absence of such a motif in the human CFTR promoter suggest that the transcriptional regulation of *CFTR* has diverged substantially since the last common ancestor of mammals and fish. Small (17 bp) sequence motifs were found to be conserved between *Fugu* and human *CFTR* intron 1 sequences. Interestingly, the conserved motifs were found to overlap a previously determined DNaseI hypersensitive site that correlates well with the transcriptional activity of the *CFTR* gene in humans (Smith *et al.,* 1995; Mogayzel and Ashlock, 2000). Short (6 bp) direct repeats and a 10 bp palindrome were identified within the conserved motifs that may play a role in the binding of regulatory proteins or forming nucleic acid secondary structures (figure 5.7). Non-coding sequence conservation was also observed between *Fugu* and human, in *CFTR* intron 9 around a candidate splice site branch point sequence (figure 5.7). This sequence was not conserved in the equivalent mouse sequence. Nucleotide sequence conservation was not observed around other reported human DNase I hypersensitive sites.

The genomic structure of *CFTR* was found to be conserved between *Fugu* and humans. Sequence similarity based analysis of the genomic sequence flanking the *CFTR* gene suggested the presence of previously undescribed protein coding genes on either side of the *CFTR* gene (section 5.4.2). DNase I hypersensitive sites both upstream and downstream of *CFTR* have been reported that do not correlate with *CFTR* expression (Smith et al 1995; Smith et al 1996; Nuthall *et al.,* 1999). These sites may instead regulate transcription of the newly identified flanking genes.

Comparisons with mouse *CFTR* genomic sequence found that the intron 1 and 9 sequences conserved between human and *Fugu* were not conserved in mouse. Similarly, a polyT tract located within intron 8 (adjacent to the exon 9 splice

acceptor site) that is polymorphic and clinically relevant in humans (Chu *et al.*, 1991; Chu *et al.*, 1993) was conserved in *Fugu* but not in mouse. These findings may relate to the known differences in mouse and human *CFTR* expression patterns and mutant phenotypes (Davidson *et al.*, 1995).

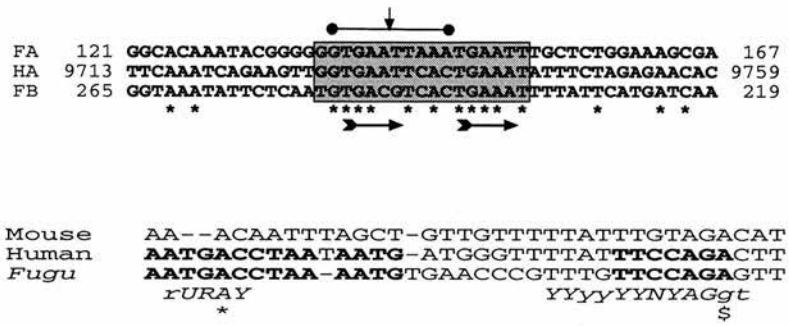


Figure 5.7; Conservation of small sequence motifs in *CFTR* introns. **(a)** Multiple sequence alignment of sequences from human (HA) and *Fugu* (FA and FB) CFTR intron 1. Coordinates indicate distance from the splice donor site of exon 1. Background shading marks the extent of the proposed, conserved sequence motif. Horizontal arrows indicate the imperfect direct repeat sequences. The horizontal dumbbell indicates the extent of the 10 bp palindrome present in the conserved sequences FB and HA, the vertical arrow indicating the axis of symmetry. Asterisks indicate 100% identity in alignment. **(b)** Multiple sequence alignment of sequence around the exon 10 splice acceptor site. Bold type indicates blocks of sequences conserved between human and *Fugu* but not in mouse. Italic sequence shows consensus canonical branch-point and splice acceptor sites (Moore, 2000). An asterisk marks the proposed branch point residue and a dollar marks the first nucleotide of *CFTR* exon 10.

5.5 Annotation anchored global sequence alignment

Subsequent to the comparative analysis of the CFTR region, a novel algorithm for genomic sequence alignment was made available (Batzoglou *et al.,* 2000). This algorithm implemented as a program called GLASS (global alignment system) used a method of iterative global alignment. The iterative alignment is performed by finding all identical sub-sequences that are in the context of a local alignment of above a threshold score. The alignments are filtered and sorted to define a set of alignment blocks that occur in the same order and orientation in both sequences being aligned. These blocks of alignment are then used as "anchors" between the sequences. The sequence between two anchors is then aligned with lower stringency than the previous alignment. This is repeated for the sequence between every pair of anchors. The result of this second round of alignments is a larger set of anchors between the sequences that are used to restrict a third and even lower stringency

alignment. This process is iterated for a predetermined number of times (eight in the original GLASS implementation). The principal of this process is summarised in figure 5.8.

The GLASS method was found to work very well for relatively homologous sequences such as human and mouse orthologous regions (Batzoglou *et al.,* 2000). Avid, a subsequent implementation of the GLASS approach (Bray *et al.,* in preparation) gave comparable results to the GLASS implementation but with substantially more efficient algorithms based on the methods of Delcher *et al.,* (1999). Avid, used in combination with a visualisation tool, Vista (Mayor *et al.,* 2000) is rapidly being accepted as a standard tool for comparative alignment and analysis of eukaryotic genomic sequence (http://pipeline.lbl.gov/; http://pga.lbl.gov/). However, for the alignment of more distantly related species, Avid was found to perform poorly, consistently failing to detect known sequence similarity between orthologous genomic sequences. The Avid alignments were found to be adversely affected by sequences of low compositional complexity that were defined as anchors early in the alignment process, preventing the subsequent biologically relevant alignment from being detected. An example of this erroneous alignment is shown in figure 5.5 (black arrow) where a protein coding exon is known to be 87% identical in alignment for 133 nucleotides but the sequence similarity is not detected by Avid. In this case, the alignment algorithm inappropriately aligned compositionally biased sequence, defining an anchor between sequences and restricting subsequent alignments. A further example of this behaviour is illustrated in figure 5.11.

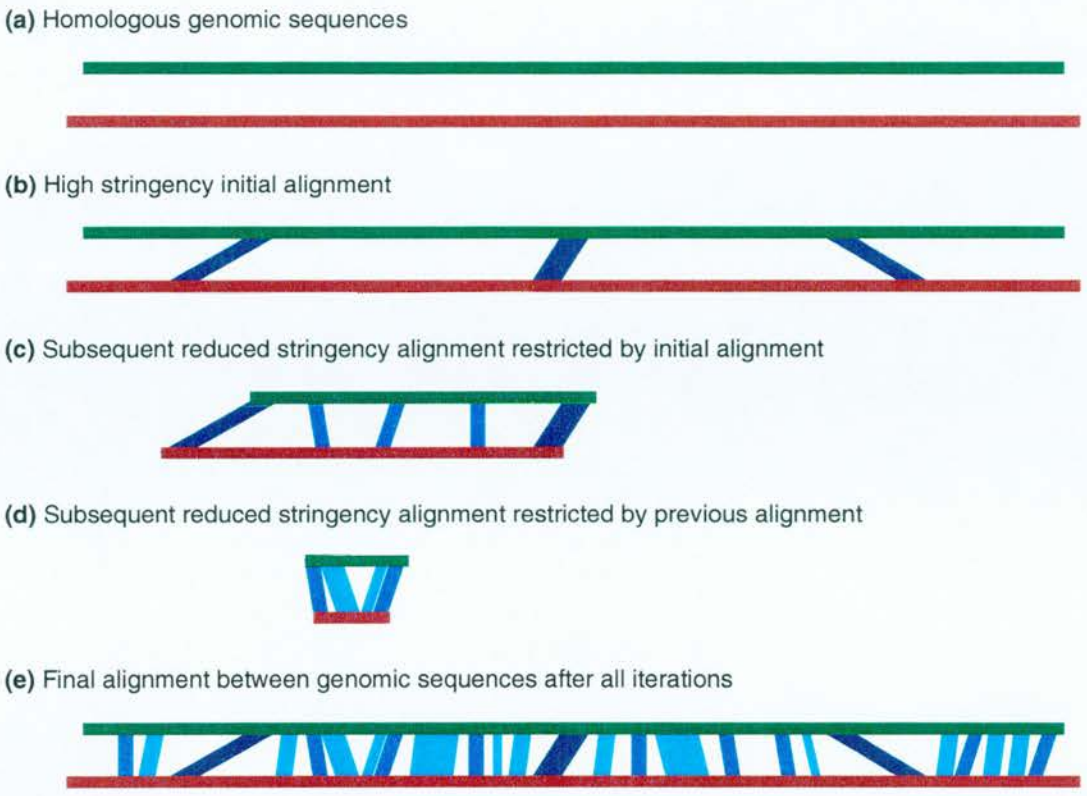


Figure 5.8; Schematic of iterative global alignment algorithm. Based on the work presented in Batzoglou *et al.*, (2000). Horizontal green and brown lines indicate non-identical, homologous genomic DNA sequences (e.g. orthologous regions of mammalian genomes). Blue lines indicate blocks of sub-sequence that align between the two parent sequences. The darker the blue, the greater the score of the alignment. The final alignment is produced through merging of all previous alignments.

5.5.1 The AAGSAL method

The methodology adopted to search for non-coding sequences conserved between *Fugu* and human *CFTR* regions, although successful was time consuming to implement and was reliant on human interpretation of dotmatrix alignments to direct further, more robust investigations. A strategy such as this would not scale well for the analysis of larger sequences and the reliance on human interpretation makes independent replication of findings problematic. However, the use of pre-defined orthologous exons to restrict the alignment was a logical approach that minimised spurious matches. In principal, using pre-defined orthologous sub-sequences (exons) to restrict dotter alignments is similar to using high confidence initial alignments to

restrict subsequent alignments (the GLASS / Avid method). The widely used PipMaker tool (Schwartz *et al.,* 2000) has a similar feature, the chaining method which acts as a post-alignment filter for local sequence alignments.

A logical extension of these methods is to use pre-defined orthologous relationships to direct the alignment. This idea was implemented as the AAGSAL method (annotation anchored global sequence alignment), the compete source code for which is included in appendix II. The implementation was in the Perl programming language that is well established as a standard language for computational molecular biology (Stein, 1996). The AAGSAL method required multiple sub-sequences to be retrieved from often large parent sequences (>1 Mb). The excessive time required by existing software (table 2.4; figure 5.9) to obtain these sub-sequences was the principal motivation to develop the sgrab algorithm (section 2.11.3). The speed advantages of this algorithm over existing software are summarised in table 2.4 and illustrated in figure 5.9.

The AAGSAL implementation takes a list of annotation based coordinates that are considered "equivalent" between the two sequences to be aligned. The between sequence equivalencies are subsequently referred to as anchors. The Avid alignment program is then used to perform the iterative global alignment as described above, restricted to aligning only between adjacent pairs of anchors. These restricted alignments are performed for all anchor pairs between the two sequences (beginning and end coordinates are considered anchors for this purpose). The results for all of the Avid alignments are then collated and a merged alignment is produced in a similar manner to the Avid and GLASS approach. Convenient annotation based coordinates that are typically used are the conserved splice sites of exons.

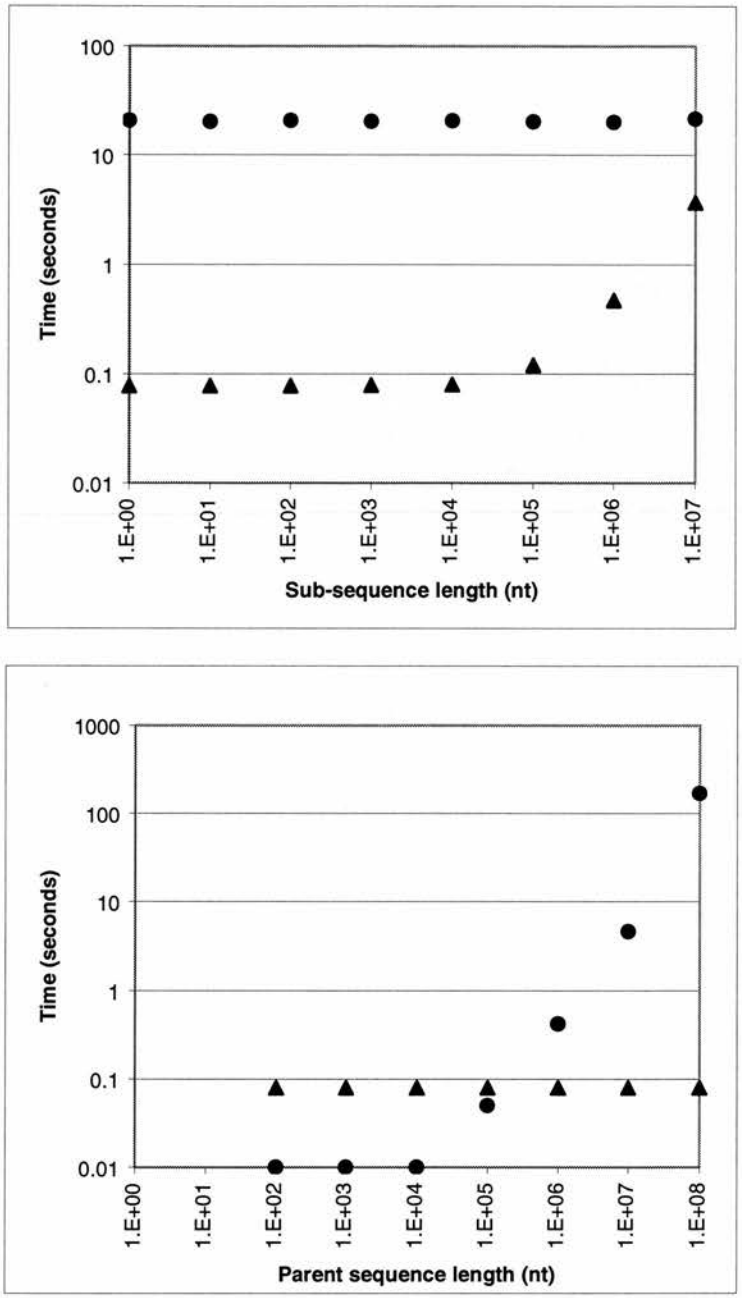


Figure 5.9; Comparison of sgrab and extract-fasta algorithms with increasing sub-sequence and parent sequence length. All values are the mean value from ten replicates of each function. Triangles indicate the sgrab algorithm and circles the extract-fasta (table 2.4) algorithm. **(a)** The real time taken to write a sub-sequence from the 47,662,662 nucleotide sequence of human chromosome 22 to a file. Sub-sequence lengths of 1 to 10,000,000 were written, the time taken measured using the Unix time function. Over the range of sub-sequence lengths tested, extract-fasta was found to take approximately the same time independent of sub-sequence length. **(b)** Real time taken to write a sub-sequence of 100 nucleotides from a parent sequence. The range of parent sequences tested was 100 to

100,000,000 nucleotides. The sgrab algorithm is unaffected by the length of the parent sequence. Sgrab is the more appropriate algorithm to use when repeatedly extracting relatively short sub-sequences from relatively long parent sequences.

5.5.2 Evaluation of the AAGSAL method

The alignments represented as Vista plots (section 2.11.2) in figure 5.10 summarise the alignment of the *CFTR* genomic sequence from Human and *Fugu* using both the standard Avid method and with the AAGSAL method described in section 5.5.1. For comparison, mouse genomic sequence is aligned to human sequence. Avid alignment of *Fugu* and human genomic sequences, in the absence of any filtering by AAGSAL did not correctly align any of the CFTR protein coding exons even though there is known to be biologically relevant nucleotide sequence conservation (section 5.4). Using the AAGSAL method described in section 5.5.1 and the coordinates of each predicted splice site as anchors between the sequences, results in the accurate alignment of 26 CFTR protein coding exons. However, as the splice sites of every exon were explicitly defined this does not represent a valid test of the AAGSAL method's ability to improve the sensitivity and specificity of Avid alignment for divergent sequences.

In a second test, using the same human and *Fugu* sequences, only the splice sites of every second exon were used as alignment anchors between the sequences. In this second test, five of the exons that were not used as alignment anchors were aligned between *Fugu* and human. This is in contrast to none when alignments were not anchored. However, nine protein coding exons remained unaligned that had sufficient sequence similarity between human and *Fugu* to produce an alignment when the exon was tightly defined by sequence anchors (figure 5.10). Further evaluation was carried out using alignment of the *Fugu* and human *TRAX* locus as a tool for evaluation (figure 5.11). Again similar results for the *TRAX* locus were found to the *CFTR* locus, the use of annotation anchors substantially aids the detection of conserved sequences in comparison to default Avid alignment but known conserved sequences remain unaligned unless they are tightly flanked by anchors (figure 5.11).

The increased ability to detect known conserved exons that are not explicitly defined to the AAGSAL method suggests that the use of *a priori* knowledge of orthologous sequences relationships can increase the sensitivity to detect previously unknown regions of sequence similarity that may be of biological significance. However, the relatively poor performance in detecting all known sequence relationships argues that the current implementation of the AAGSAL – Avid alignment method could be substantially improved. An obvious starting point to improve the performance is to modify the thresholds used by Avid to define sequence anchors. Currently these thresholds are optimised for human – mouse genomic sequence comparison. As these thresholds are "hard coded" into the Avid software they cannot be modified without access to the source code, which has not yet been released. Other interesting ways of developing this method could be to combine annotation anchoring with the chaining method of PipMaker (section 2.11.2) or the "DNA block aligner" (dba, a component of the Wise2 package section 2.11.2).

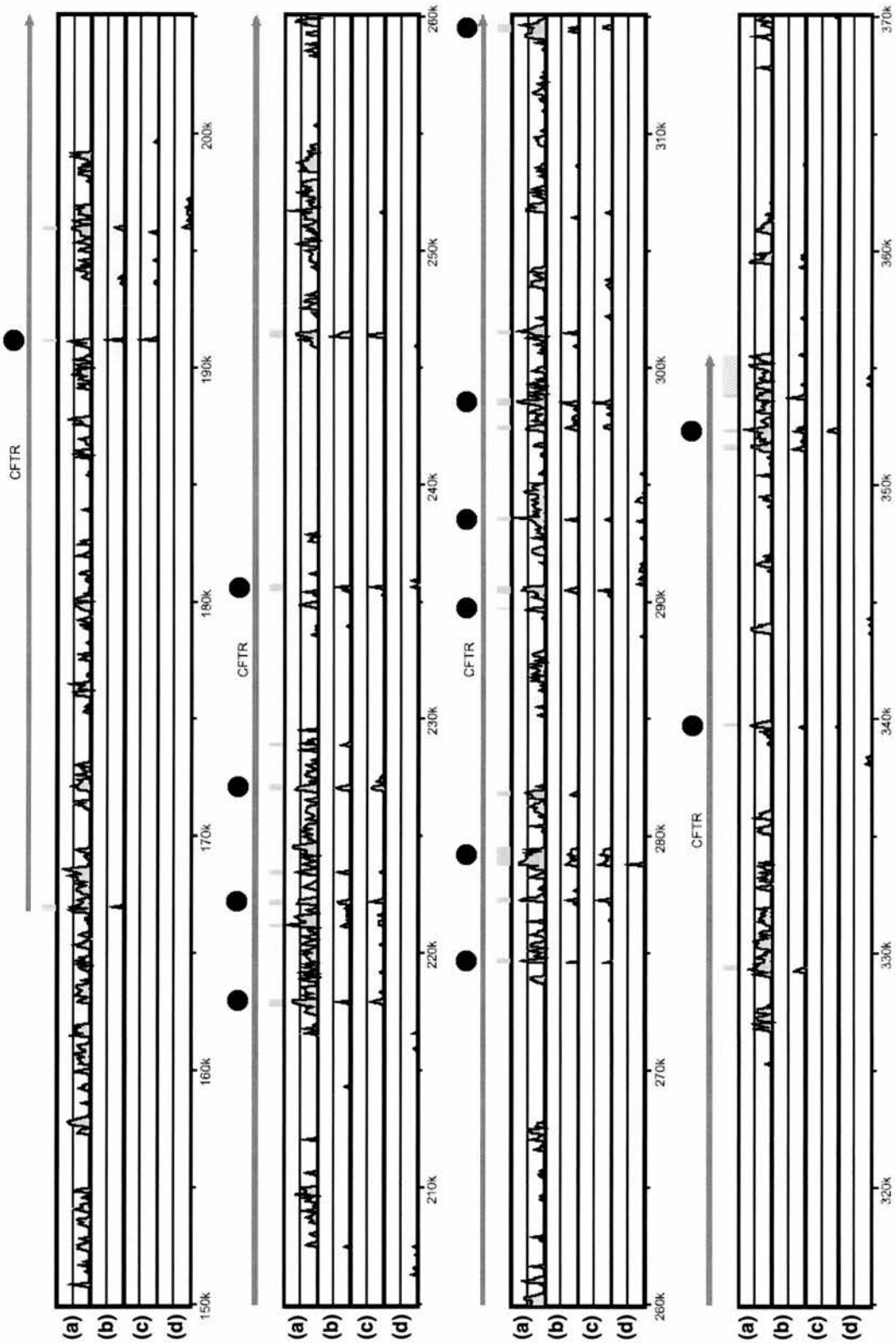

CFTR
(a)
(b)
(c)
(d)
150k
160k
170k
180k
190k
200k
210k
220k
230k
240k
250k
260k
270k
280k
290k
300k
310k
320k
330k
340k
350k
360k
370k

Figure 5.10; Evaluation of the AAGSAL method – the *CFTR* locus. Vista plot (section 2.11.2) of Avid alignments between *Fugu*, human and mouse *CFTR* genomic regions. **(a)** Avid alignment between human and mouse *CFTR* genomic sequences highlighting the conservation of protein coding exons all of which are correctly aligned between human and mouse. **(b)** Avid alignment between human and *Fugu* sequences restricted by the AAGSAL method using known and predicted splice donor and acceptor sites as anchors between sequences. **(c)** Avid alignment between human and *Fugu* sequences restricted by the AAGSAL method using known and predicted splice donor sites as anchors between sequences. Only the splice sites of exons marked with a black circle were used as anchors. **(d)** Default avid alignment between human and *Fugu* sequences.

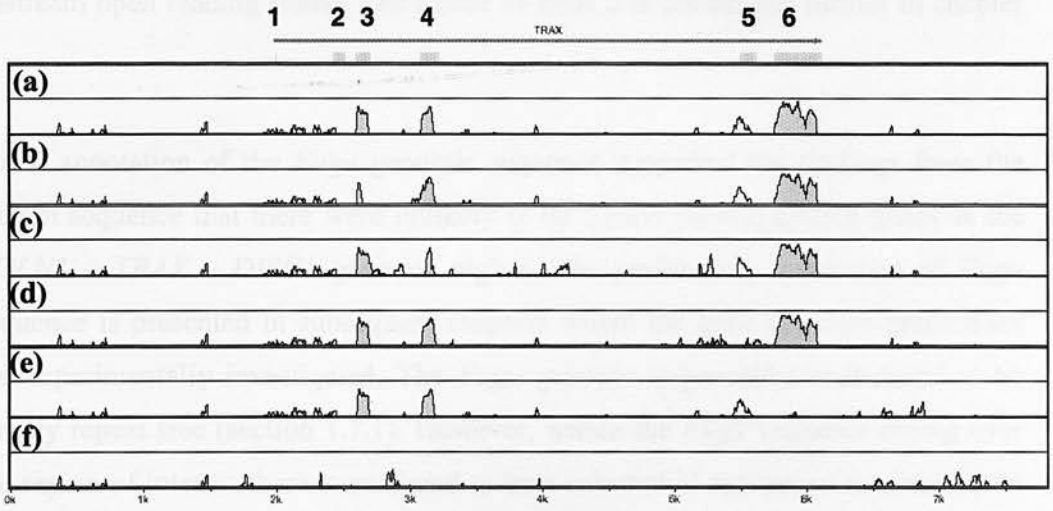


Figure 5.11; Evaluation of the AAGSAL method – the *TRAX* locus. Vista plot (section 2.11.2) of Avid alignments using a range of annotation anchors to restrict the sequence alignment. Alignments are between the *Fugu TRAX* locus (chapter 8) and the human *TRAX* locus. The coordinates along the horizontal axis relate to Fugu genomic sequence. The relative position of protein coding exons is indicated above the graph and the transcriptional orientation of TRAX indicated by a horizontal arrow. The graphs show 50% to 100% sequence identity in alignment averaged over a 60 nucleotide sliding window. The horizontal line at the mid point of each graph indicates 75% sequence identity. **(a)** Avid alignment restricted by the AAGSAL method using start and stop codons and the splice sites of each TRAX exon as annotation anchors. **(b)** The splice sites of exon 3 were not used as annotation anchors. **(c)** The splice sites of exon 4 were not used as annotation anchors. **(d)** The splice sites of exon 5 were not used as annotation anchors. **(e)** The splice sites of exon 6 were not used as annotation anchors. **(f)** Avid alignment without AAGSAL restriction of the alignment.

5.6 Discussion

Preliminary annotation of human genomic sequence identified three processed pseudogenes and two spliced and polyadenylated transcripts of unknown protein coding potential. The protein coding status and functional relevance of these transcripts (*Backtrax* and *Foretrax*) is considered later, in comparison to the orthologous *Fugu* sequence (section 8.5.7). Several clusters of ESTs and single ESTs were identified within the introns of the *DISC1* gene (summarised in section 6.2.1). Investigation of the *DISC1* intragenic EST clusters is reported in chapter 6. No evidence was found to suggest that the *DISC2* transcript was associated with an upstream open reading frame. The nature of *DISC2* is considered further in chapter 7.

Initial annotation of the *Fugu* genomic sequence supported the findings from the human sequence that there were unlikely to be further protein coding genes in the *EGLN1 – TRAX – DISC1* genomic region. The preliminary annotation of *Fugu* sequence is presented in subsequent chapters where the gene structure predictions are experimentally investigated. The *Fugu* genome is generally considered to be largely repeat free (section 1.7.1). However, within the *Fugu* sequence contig over the region of interest there were found to be a substantial number of tandem repeat arrays that is not typical of gene containing regions of the *Fugu* genome. These repeats are not related to each other in the primary sequence of the basic repeat unit (table 5.1) and they do not represent a conserved feature of the locus (section 5.3.1). The sequence of overlapping cosmid clones also suggested that these repeats may be of polymorphic length in the *Fugu* population. It has previously been proposed (Elgar *et al.,* 1996) that the majority of repeats in the *Fugu* genome are clustered into specific regions such as near the telomeres and centromeres. The finding of interspersed repetitive elements and the high number of tandem repeats, particularly towards the 3' end of *DISC1* suggest that the *DISC1* locus is at or near the boundary of a repeat rich region of the *Fugu* genome. If this is the case, it may relate to the lack of cosmid or BAC clone library coverage distal to the *DISC1* gene (section 3.4.3).

Using the *CFTR* locus as a model for human to *Fugu* genomic alignment, optimised parameters for PipMaker were developed. The idea of using pre-defined orthologous sequence relationships to restrict sequence alignments was formulated. Application of this approach to the *CFTR* locus identified conserved sequences in introns 1 and 9. A fully automated implementation of annotation anchored global sequence alignment (AAGSAL) was written.

The AAGSAL method is essentially a pre and post filter for Avid alignment that allows the incorporation of *a priori* biological knowledge to direct the early stages of alignment. Such intervention is generally unnecessary for the alignment of relatively similar sequences such as orthologous human and mouse loci, but substantially aids the alignment of more divergent sequences such as fish to mammal. An interesting side effect of the AAGSAL method is the reduction of total memory requirement for the alignment of two sequences, allowing the alignments to be carried out on machines with less memory. AAGSAL can also reduce the total computational load for the alignment of two sequences (compared to Avid) although this is dependant on the length of sequences and spacing of anchors. The AAGSAL method was demonstrated to greatly increase the sensitivity and specificity of Avid alignment when aligning *Fugu* and human genomic sequences. However, in a subset of cases known sequence similarity was still not accurately aligned using the AAGSAL method unless the sequences were tightly flanked by anchors. It is suspected that this system can be improved by the optimisation of Avid parameters.

Chapter 6

Sequence analysis of *DISC1*

6.1 Preface

DISC1 is the strongest positional candidate for a contribution to mental illness susceptibility in the t(1;11) translocation (section 1.6.6). However, in the absence of recognised protein domains or homologous sequences little could be reliably predicted of the function of the gene. The existence of EST clusters located within the introns of *DISC1* also indicated that there may be additional, previously unknown *DISC1* splice variants. This chapter describes the *DISC1* gene in humans, refining the previously determined transcription map. Homologues of *DISC1* are also identified in a range of vertebrates providing a means of investigating evolutionary and by inference functional constraints on gene structure, splicing and amino acid sequence.

6.2 The genomic structure of human *DISC1*

Human *DISC1* was reported to consist of 13 exons, with a primary transcript of 6913 nucleotides and a common alternative splicing event utilising different exon 11 splice donor sites (Millar *et al.,* 2000a). The alternative exon 11 donor site usage lead to the differential inclusion of 66 nucleotides and maintenance of the *DISC1* reading frame. Intron 12 was reported to be a rare AT-AC type intron (section 6.5.3 and Millar *et al.,* 2000). Northern blot analysis demonstrated a major transcript of 7.5 kb detected in all adult tissues tested, with additional smaller bands of 1.4 kb and 1 kb which were thought to be non-specific artefacts that had been observed previously using non-*DISC1* probes (Millar *et al.,* 2000a and K Millar, personal communication). The initial transcript map of the *DISC1* region (figure 6.1) was based on EST identity to cosmid clone end sequences and hybridisation onto the cosmid contig of ESTs that had previously been mapped (http://www.ncbi.nlm.nih.gov/genemap/) to the 1q42 region (K. Millar and S. Christie, personal communication). From the initial

transcription map, it was clear that multiple ESTs originated from the *DISC1* locus but did not represent the previously described *DISC1* or *DISC2* transcripts (figure 6.1). Northern blot and RT-PCR analysis were carried out to investigate the nature and significance of these transcribed sequences.

6.2.1 Refining the transcription map of human *DISC1*

To confirm the mapping of ESTs to the clone contig (figure 6.1), PCR was carried out on contig clones. Genomic DNA positive controls were included to detect instances where primers flanked a splice site resulting in a false negative PCR mapping result. With the exception of EST AA361299, all of the ESTs of the initial transcript map (figure 6.1) were found to map to the expected clones in the cosmid / PAC contig. Only cosmid clone J0942 was positive by PCR for EST AA361299. As overlapping clones I0142 and PAC 27-B9 were PCR negative, cosmid J0942 was removed from the contig under suspicion of being a chimeric clone. It was subsequently shown that EST AA361299 maps to chromosome 1q23.3 (tens of megabases from the breakpoint locus) on the basis of the draft human genome assembly (The International Human Genome Sequencing Consortium, 2001) and the accession maps project (section 2.11.1). ESTs AA361879, AA885025, HS833289, HSZZ52663, AA610789 and the ESTs representing unigene cluster Hs.212335 were investigated by RT-PCR and Northern analysis. All of the tested ESTs were detected by RT-PCR in foetal heart and brain tissue (table 6.1), but not by Northern analysis (data not shown) while an actin control probe gave clear signals (figure 6.2).

Multiple pairwise RT-PCR reactions were carried out, attempting to link ESTs and EST clusters with other ESTs, *DISC1* exons or the *DISC2* transcript (table 6.1). Reactions were not carried out in a random manner, rather they were ranked based on the proximity of features and inferred orientation of ESTs from directionally cloned libraries. For example EST AA361879 represented a 5' end read of a directionally cloned cDNA and could be oriented relative to the end sequence of cosmid C2125 with which the EST was originally identified. An uninterrupted reading frame and inferred co-linear transcriptional orientation with *DISC1* lead to the subsequent hypothesis that AA361879 may represent an alternative exon of *DISC1*. RT-PCR

demonstrated that AA361879 was in a transcript with *DISC1* exon 3, but that the EST represented non-coding sequence in the 3' UTR of a short *DISC1* isoform.

Using RT-PCR to link transcribed sequences was only informative if a positive result was obtained. Positive results are summarised in figure 6.3. A negative result could indicate that the expressed sequences were not part of the same transcript, that the primer annealing sites were physically too far separated to produce a PCR product or they were not in appropriate relative orientations to produce a product. None of these PCR reactions were optimisable for temperature or salt concentration due to a lack of templates for positive control. Consequently, a negative result may also have been a false negative due to sub-optimal PCR conditions.

Products of the RT-PCR reactions were sequenced to confirm they represented the expected sequence at both ends and define the splices sites through alignment to genomic sequence. The assembly of human genomic sequence over the *DISC1* region (section 4.3) allowed the RT-PCR product sequences to be aligned with the genomic sequence, ESTs and mRNA sequences to produce a single base resolution transcription map over the *DISC1* genomic region incorporated in to the DISCACE database (section 5.2).

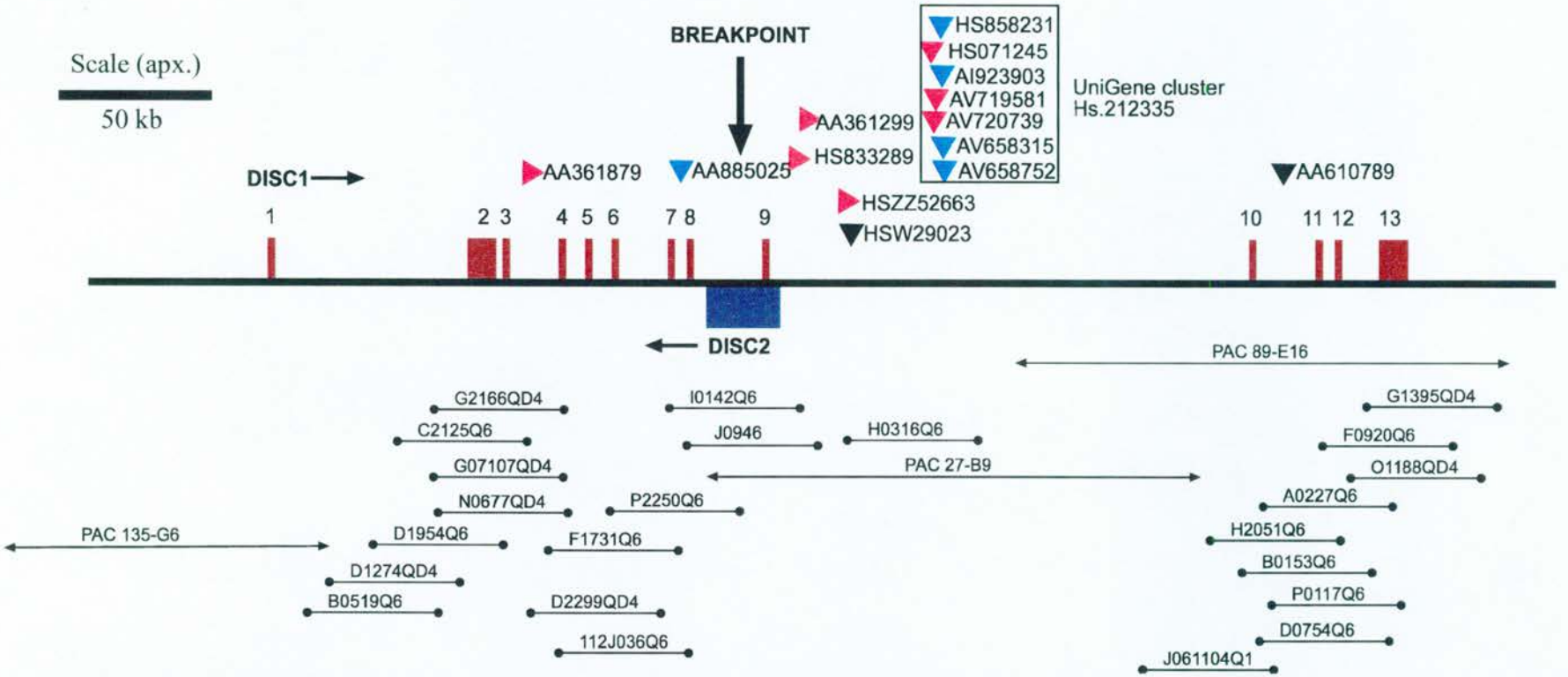


Figure 6.1; Initial transcript map of the human *DISC1* region. Cosmid clones identified across the *DISC1* region are indicated by horizontal dumbbells, PAC clones are represented by double headed arrows. The relative position of exons (rectangles, red for *DISC1*, Green for *DISC2*) and EST sequences were determined by hybridisation onto the cosmid and PAC clones, or through alignment to the sequenced ends of each clone insert. The relative transcriptional orientation of *DISC1* and *DISC2* are indicated by black arrows. ESTs derived from cDNA clones without directional information are indicated by black triangles, those representing the 5' end of cDNA clones are coloured red and those representing 3' ends in blue. The label "BREAKPOINT" indicates the position of the chromosomal breakpoint. Exons are not drawn to scale. Where directional information on transcriptional orientation could be inferred through sequence alignment triangles point the direction of transcription. Figure adapted from Millar, personal communication with permission.

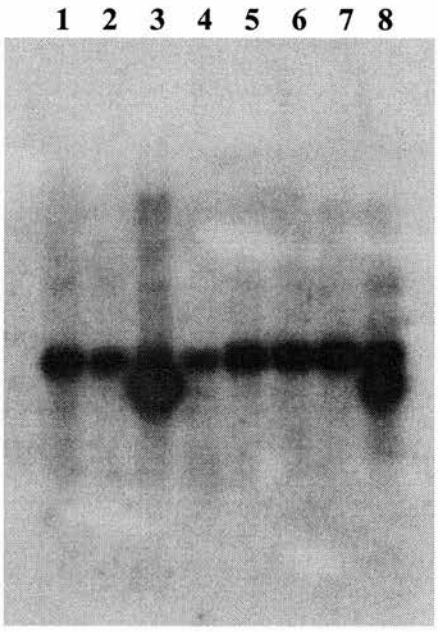


Figure 6.2; Control Northern hybridisation. Clontech multiple tissue Northern blot (section 2.9.7) hybridised with an actin control probe. Lanes 1 to 8 are feotal human poly(A) enriched RNA from pancreas, kidney, skeletal muscle, liver, lung, placenta, brain and heart respectively. The horizontal arrow indicates the position of the 2.4 kb size marker on the Northern blot filter. The clear signal indicates that the Northern blot and the technique used work. The absence of signal when the EST sequences summarised in table 6.1 were used, therefore indicating a low level or absence of expression of the EST sequences in these tissues.

EST	RT-PCR		Northern	RT-PCR link
EMBL acc.	Brain	Heart		
AA361879	+	+	-	Exon 3 of *DISC1*, represents an alternative 3' UTR.
AA885025	+	+	-	-
HS833289	+	+	-	Exon 9 of *DISC1*, represents an alternative 3' UTR.
HSZZ52663	+	+	-	-
HSW29023	+	+	-	-
Hs.212335	+	+	-	Spliced with exon 9 of *DISC1*, represents an alternative terminal exon and 3' UTR of *DISC1*.
AA610789	+	+	-	-

Table 6.1; Summary of RT-PCR and Northern analysis for 'anonymous' ESTs from the *DISC1* transcription map. In RT-PCR columns, '+' indicates a product of the expected size was observed in the RT-PCR reaction and no product was observed in negative controls. For each of the RT-PCR reactions summarised, a genomic DNA positive control was used and resulted in a single product of the same size as the RT-PCR product. 'Northern' indicates

results of Northern blot hybridisation with radioactively labelled probes designed from the sequence of each EST. Clontech multiple tissue Northern blots (section 2.9.7) were used for the hybridisation, no hybridisation signal was detected for any of the probes used. 'RT-PCR link' indicates positive results from the multiple pair-wise RT-PCR reactions between ESTs and known transcripts.

6.2.2 Alternate human *DISC1* splicing

The alternate inclusion or exclusion of exon 11 distal sequence resulted in two splice variants with a total mature transcript length of approximately 7.5 kb (Millar *et al.*, 2000a). The transcript including the distal portion of exon 11 had an open reading frame of 854 codons and is subsequently referred to as the L1 isoform (Long one). The transcript missing the distal region of exon 11 had an open reading frame of 832 codons and is subsequently referred to as the L2 isoform (Long two).

A further three alternative *DISC1* transcripts were identified through EST alignment to genomic sequence and RT-PCR (table 6.1). An isoform was identified that spliced from the exon 9 splice donor site to a previously unidentified exon (subsequently exon 9a). The open reading frame continued for only one codon before an in-frame stop codon in exon 9a. There were subsequent stop codons in all three forwards reading frames. This splice variant was first suggested by the Incyte® EST sequence 2320731r6 and was subsequently confirmed by RT-PCR using human foetal and adult heart cDNA. If exons 1 to 9 were included in the transcripts containing the terminal exon 9a, this splice variant would have an open reading frame of 662 codons. This splice variant is subsequently referred to as the S1 isoform (Short one).

A second distinct transcript also resulted in the truncation of the *DISC1* reading frame shortly after the exon 9 sequence. This S2 isoform (Short two) was first identified by RT-PCR between the EST HS833289 and exon 9 of *DISC1*. The RT-PCR product was found to be unspliced and represented a failure to utilise the exon 9 splice donor site. From this RT-PCR it was not possible to determine if the transcript detected represented *DISC1* or *DISC2* although the transcriptional orientation inferred from the directionally cloned EST HS833289 was consistent with it representing a *DISC1* transcript.

A splice variant of *DISC1* was also identified by RT-PCR (table 6.1) that did not splice from the exon 3 splice donor site, resulting in a truncation of the *DISC1* reading frame two codons after the exon 3 splice donor site. Subsequently, the full insert sequence of a cDNA clone was submitted to the EMBL database (section 2.11.1) that represented this RNA processing event (EMBL:AK025293). However, the cDNA sequence AK025293 appeared to utilise an alternative exon 1. Use of the alternative exon 1 could not be confirmed by RT-PCR (data not shown and personal communication, R. James). The cloning of a mouse *DISC1* transcript representing the 'known' exon 1, exon 2, exon 3 and the subsequent failure to splice using the exon 3 splice donor site (R. Devon, personal communication) suggests that this represented an evolutionarily conserved splice variant of *DISC1*. This isoform is subsequently referred to as the E3 (exon 3 truncation, or extremely short) isoform.

Each of the *DISC1* isoforms discussed are summarised in figure 6.3. By Northern blot, only a 7.5 kb transcript was detected for *DISC1* in a range of adult and embryonic tissues (Millar *et al.,* 2000). It remains unclear if the observed band represents the L1 or L2, or both L1 and L2 splice variants of *DISC1*. Using constitutive *DISC1* exons as probes for Northern blot hybridisation did not detect the expression of S1, S2 or E3 isoforms. Using probes specific to each of the distinct 3' UTRs of these three isoforms also failed to detect expression in Northern blot hybridisation (table 6.1). This suggested that S1, S2 and E3 represent rare transcripts or their expression is temporally or spatially restricted and appropriate tissues were not represented on the Northern blots used (figure 6.2). The S2 isoform in particular was represented by multiple ESTs (figure 6.1) suggesting its expression at appreciable levels. There was no obvious bias in the distribution of tissue of origin for the ESTs representing the S2 isoform.

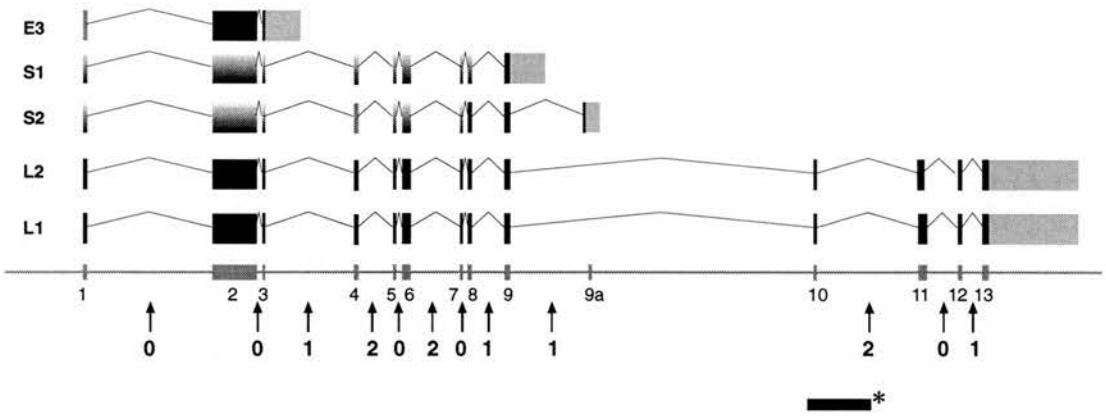


Figure 6.3; Summary of alternate human *DISC1* splicing. The grey horizontal bar represents human genomic sequence, the relative location of each exon is shown to scale (* indicates 30 kb). For each of the spliced transcripts (E3, S1, S2, L1 and L2) light grey boxes indicate 3' UTR. Black boxes indicate protein coding exons that have been experimentally demonstrated to be in a particular transcript. Boxes with a gradient fill indicate exons that are assumed to be in the transcript but no single cDNA clone containing all of the transcript has been isolated. Numbered arrows indicate the splice phase of each intron.

6.3 The genomic structure of *Fugu DISC1*

6.3.1 Gene structure prediction

The genomic structure of *DISC1* was poorly predicted by *ab initio* methods in both human and *Fugu* sequences (section 5.2 and figure 6.4 respectively). Human *DISC1* has previously been reported to consist of 13 exons, all contributing to the 854 codon open reading frame (Millar *et al.,* 2000a). Manually integrating *ab initio* gene prediction with homology to human DISC1 amino acid sequence (TBLASTN and Genewise), predicted *Fugu* orthologues of exons 2, 3, 4, 7, 8, 9, 10, 11, 12 and 13 (figure 6.4).

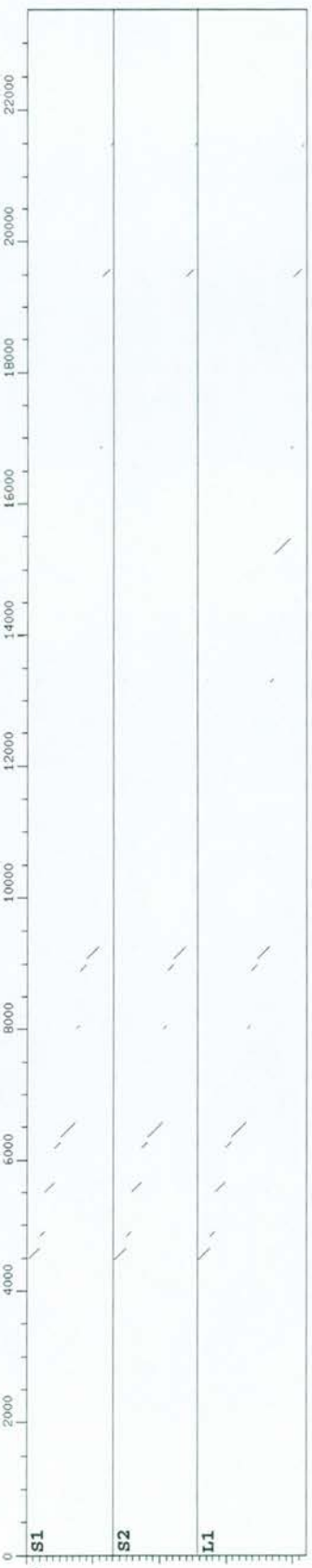


Figure 6.4; Gene structure prediction of *Fugu DISC1*. The region of *Fugu* sequence shown corresponds to nucleotides 12501 to 36000 of the *Fugu* sequence contig (section 3.5.6). **(a)** Dotter (section 2.11.2) alignment of *Fugu DISC1* cDNA sequences plotted against the *Fugu* genomic sequence contig (vertical and horizontal axis respectively). The sequences 'S1', 'S2' and 'L1' are cDNA sequences representing alternatively spliced forms of the *Fugu DISC1* transcript (section 6.3.2). A dotter window size of 35 was used for this alignment. **(b – g)** *Ab initio* gene prediction. Exon predictions are indicated by black boxes, only forward strand predictions are shown: **(b)** Genefinder, **(c)** Genscan, **(d)** Fgenes, **(e)** Genemark, **(f)** Hmmgene, **(g)** Grail2. Red boxes indicate homology based gene prediction: **(h)** TBLASTX of *Fugu* genomic sequence versus the assembled human *DISC1* genomic sequence (section 4.3), **(i)** Genewise using human DISC1 amino acid (L1 isoform) as the query sequence.

The standard settings and references for each program are summarised in section 2.11.2. **(j)** Manual prediction of *Fugu DISC1* gene structure integrating the data from 'a' to 'i' as well as considering the structure of human *DISC1*. **(k)** The gene structure of *Fugu DISC1* (yellow boxes) based on alignment of cDNA to genomic sequence (exon 1 is homology and gene model based prediction). The numbers associated with each exon reflects the *DISC1* nomenclature used in all subsequent sections.

6.3.2 Demonstrating the genomic structure of *Fugu DISC1*

The eleven *Fugu* cDNA libraries detailed in section 2.9.7 were screened by high stringency hybridisation using DNA probes corresponding to *Fugu* exons 2 and 8 – 9. No cDNA clones representing transcription of the *DISC1* gene were identified (data not shown). RT-PCR on *Fugu* heart and ovary RNA (the only RNA samples available) between predicted exons 2 and 13 produced 2 products of ~1.4 kb and ~1.6 kb (figure 6.5). The 1.4 kb and 1.6 kb RT-PCR products were isolated and cloned for further analysis.

Twenty clones of the 1.4 kb RT-PCR product and 20 clones of the 1.6 kb product were analysed by restriction digestion (figure 6.5) and sequencing. Of the 1.6 kb product derived clones, 19 represented the same insert and one clone contained an insert of 1.5 kb that was not related to *DISC1*. The 20 1.4 kb product derived clones represented two inserts differing in size by 45 bp (figure 6.5). Eight of these clones had the additional 45 bp, the remaining twelve did not. The 1.6 kb product and both 1.4 kb products represented alternatively spliced forms of *DISC1*. Alignment of the cDNA clone sequences to the *Fugu* genomic sequence contig demonstrated the genomic structure of *Fugu DISC1* (figure 6.4). The 45 bp sequence differentially included in the 1.4 kb products was also present in the 1.6 kb product. Through alignment to genomic sequence, this 45 bp sequence was shown to represent an additional exon between orthologues of human exons 11 and 12. This exon is subsequently referred to as exon 11a.

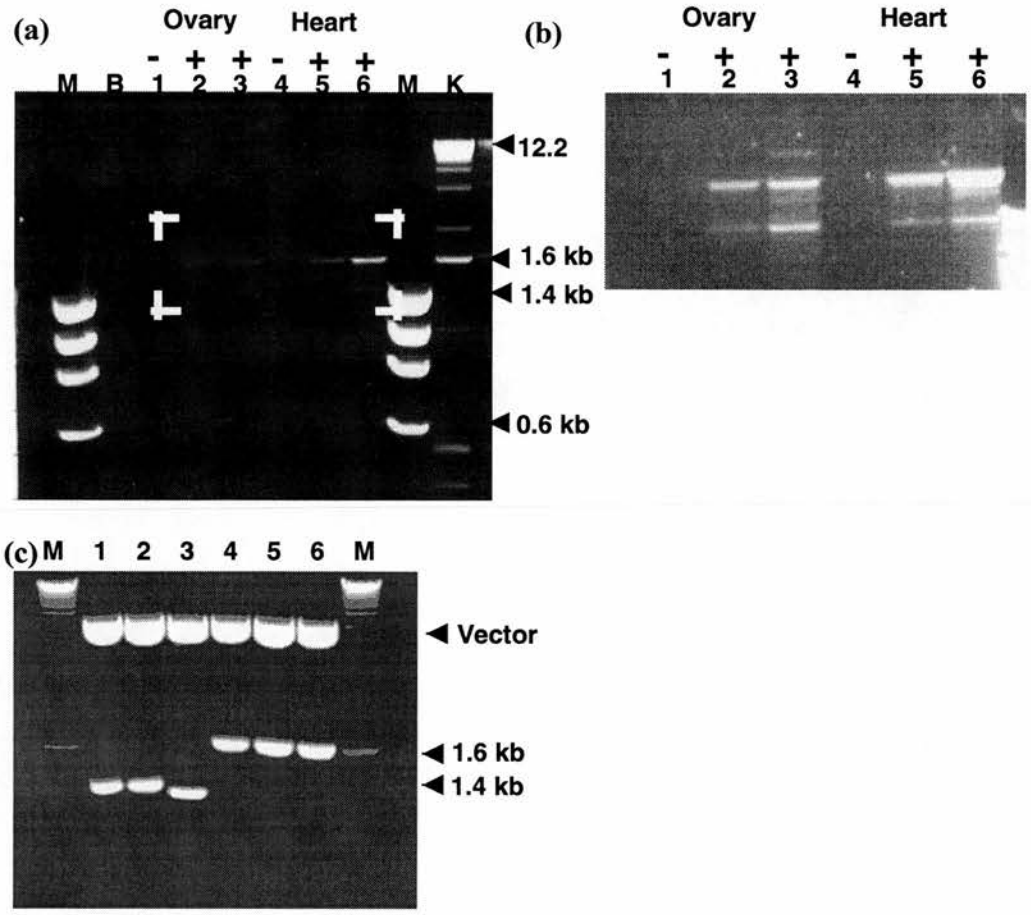


Figure 6.5; Expression and alternative splicing of *Fugu DISC1*. **(a)** RT-PCR between predicted *Fugu DISC1* exons 2 and 13. Lanes labelled 'M' and 'K' indicate molecular weight size markers ΦX174 *HaeIII* and 1 kb ladder respectively (section 2.7.2). 'B' indicates distilled water was used as a template for negative control. Lanes 1, 2 and 3 represent reactions that used *Fugu* ovary RNA as a template. Lanes 4, 5 and 6 represent reactions that used *Fugu* heart cDNA as a template. Lanes 1 and 4 were RT-PCR negative controls where distilled water was used instead of reverse transcriptase during the reverse transcription step. Reverse transcription reactions in lanes 2 and 5 were primed with random oligonucleotides. Reverse transcription reactions in lanes 3 and 6 were primed with hexa-deoxy-thymine oligonucleotides. The region between crosshairs is shown in greater detail and longer exposure time in panel 'b'. **(b)** The region of specific interest from panel 'a' in greater detail. Lane numbering is the same as described for panel 'a'. **(c)** RT-PCR product sub-clones restriction digested with *EcoRI* to reveal the insert size. Individual bands were cut from the gel shown in panels 'a' and 'b' and sub-cloned. Lanes marked 'M' indicate the 1 kb ladder DNA molecular weight maker. Lanes 1 – 3 were sub-clones of the lower 1.4 kb band in panel 'b'. Lanes 4 – 6 were sub-clones of the larger 1.6 kb band in panel 'b'. Note the slightly

smaller size of sub-clone insert in lane 3, demonstrating that the 1.4 kb RT-PCR band represents a doublet although it was not resolved on the gels shown.

6.4 *DISC1* in other model organisms

6.4.1 The genomic structure of *Tetraodon DISC1*

Ordered and oriented sequence contigs from the targeted assembly of *Tetraodon* whole genome shotgun sequence data (section 5.3.2) were used to predict the genomic structure of *Tetraodon DISC1*. Alignment of the *Fugu DISC1* cDNA sequence with the *Tetraodon* sequence identified regions with high sequence identity (>70%) corresponding to each of the *Fugu DISC1* exons with the exception of exon 10 (figure 6.11). The absence of exon 10 homology was coincident with a gap between sequence contigs. Consequently, the observed absence of exon 10 most likely reflected a lack of sequence coverage rather than the evolutionary loss of the exon. The conservation of exonic sequence and splice site consensus sequences (section 5.3.2 and data not shown) suggested that *Tetraodon DISC1* had an identical genomic structure to *Fugu DISC1*. Of particular note was the conservation of exon 11a, an exon not identified in human *DISC1*.

6.4.2 A zebrafish homologue of *DISC1*

Searching the EST subset of the EMBL database (section 2.11.1) with *Fugu DISC1* amino acid sequence identified a single *Danio rerio* (zebrafish) EST (EMBL:BG799396) with 59% identity in translation to exon 13 of *Fugu DISC1*. Searching the zebrafish whole genome shotgun data set (section 2.11.1) with *Fugu DISC1* amino acid sequence as a query for TBLASTN identified two overlapping shotgun sequences. The shotgun sequences were 59% identical in translation to the 3' end of *Fugu DISC1* exon 2. Iterative assembly of zebrafish whole genome shotgun sequences identified a further three overlapping sequences that were used to extend the genomic contig and improve consensus sequence quality (figure 6.6).

Primers were designed to the predicted zebrafish *DISC1* exons 2 and 13 sequences. Using zebrafish 36 hour whole embryo cDNA (a gift from L. Haines, MRC Human

Genetics Unit, Edinburgh) as a template for RT-PCR between the predicted exons 2 and 13 produced a 1.4 kb PCR product (figure 6.6). Cloning and sequencing (sequencing was carried out by J. Fantes) of the PCR product confirmed that a zebrafish homologue of *DISC1* had been identified and isolated (figure 6.13). The zebrafish RT-PCR reaction only produced one abundant product, where as the equivalent *Fugu* reaction between exons 2 and 13 had produced three products, each representing an alternatively spliced form of the *DISC1* transcript (section 6.3). However, tissue and developmental stage was not matched between species excluding meaningful comparisons. It is interesting to note that the zebrafish *DISC1* exon 2 sequence identified using the *Fugu* DISC1 amino acid sequence was not detected by the same methods when using mammalian DISC1 amino acid sequence.

The *DISC1* open reading frame of exon 2 was observed to be uninterrupted by in-frame stop codons upstream of the demonstrated exon 2 sequence (figure 6.7). Combined with the known genomic structure of *DISC1* in human, mouse, *Tetraodon* and *Fugu* this suggested that zebrafish exon 2 extends beyond the assembled sequence. Interestingly, the continuing open reading frame has a region of expanded tandem repeat suggesting that zebrafish *DISC1* may have a larger N-terminal region than the other homologues characterised to date. This possibility and its implications are discussed in section 6.7.

The cloning of a zebrafish homologue of *DISC1* was notable for two reasons. Firstly, it provided an additional diverged sequence for the analysis of DISC1 amino acid conservation (section 6.7). Secondly, the failure to detect *DISC1* in mouse embryos by *in situ* hybridisation (J. Fantes, personal communication) required the selection of an alternate model organism for the investigation of *DISC1* expression, particularly during development. The zebrafish is well established as a model organism for investigating gene expression during development (Wixon, 2000 for review), and is an excellent model for studying the function of genes in vertebrate development by the rapid morpholino knock down method (Ekker, 2000 for review). Sequence of the zebrafish *DISC1* transcript is a necessary pre-requisite for both of these strategies.

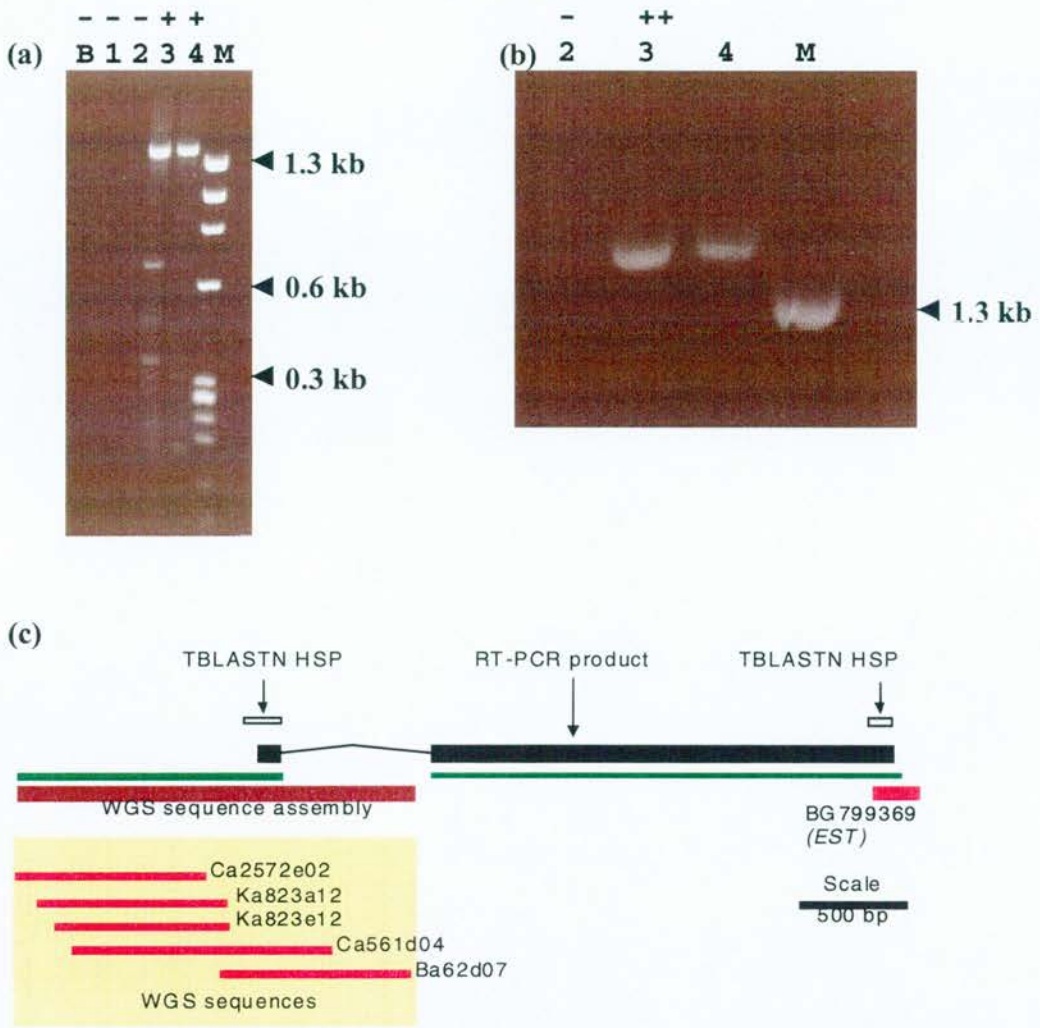


Figure 6.6; A zebrafish homologue of *DISC1*. **(a)** RT-PCR products of zebrafish *DISC1* (exons 2 – 13) amplified from 36 hour whole embryo cDNA. cDNA synthesis was primed with random oligonucleotides (lane 3) or oligo(dT) oligonucleotides (lane4). Lanes 1 and 2 were reverse transcription negative control reactions (section 2.6.3) and Lane B was a PCR negative control reaction. Lane M contains the ϕX174 *HaeIII* DNA molecular weight marker. **(b)** Enlarged view of the ~1.4 kb RT-PCR products after the products had been further separated by electrophoresis to resolve any 'doublets' (two bands of a similar size that are not readily separated by electrophoresis). No doublets were observed. **(c)** Strategy for cloning zebrafish *DISC1*. TBLASTN searching of *Fugu* DISC1 amino acid sequence against zebrafish EST sequences and whole genome shotgun traces identified two regions of sequence similarity (open boxes). The zebrafish EST BG799369 (thick horizontal red line) showed sequence similarity (52% identity over 127 amino acids) to the C-terminus of DISC1. The identified whole genome shotgun traces (Ca561d04 and Ba62d07) were used to seed the iteratively assembled process (section 2.11.4) extending and improving the quality of

sequence. RT-PCR between the two blocks of *DISC1* homologous sequence (panels 'a' and 'b') demonstrated the co-occurrence of these homologous sequences in a single transcript (black rectangles). The carat shaped line between black rectangles indicates a splice site. The extent of uninterrupted reading frames is indicated by horizontal green bars. Note the extended open reading frame in genomic sequence upstream of the demonstrated transcript (see section 6.7).

6.4.3 The genomic structure of mouse *DISC1*

Sequence representing mouse *DISC1* exons 1 to 8 was previously generated (R. Devon, personal communication). Using the available mouse exon 1 to 8 sequence and TBLASTN searching with human DISC1 amino acid sequence, homologous sequences were identified in mouse whole genome shotgun sequence data (Public Mouse Sequencing Consortium). The mouse *DISC1* homologous sequences were used to seed an iterative assembly process (section 2.4) with a modification that utilised trace quality data as well as raw sequence (section 2.4). Assemblies of mouse whole genome shotgun sequence were produced representing exons 1, 2, 7, 8, 9, 10, 11, 12 and 13 of *DISC1* and the associated flanking intronic sequence.

For every exon of human *DISC1*, a maximum of one mouse sequence assembly was produced, suggesting that there is only one *DISC1* homologous (TBLASTN bit score of >55 using human DISC1 amino acid sequence) sequence in the mouse genome. This finding is in agreement with nucleotide hybridisation experiments using regions of human *DISC1* coding sequence to probe mouse whole genomic DNA Southern blots (R. Devon, personal communication). The single copy nature of *DISC1* in the mouse genome allowed the mouse contigs to be ordered and orientated based on the linear organisation of exons in the human genomic sequence contig (figure 6.6). This ordering of mouse contigs was further validated by nucleic acid hybridisation and by PCR using the mouse BAC contig (section 4.4) as well as sequence overlaps with BAC clone end sequences (figure 4.5).

Combining the sequence data from exons 1 to 8 and the clustered whole genome shotgun data provides a prediction of the full open reading frame, of the longest isoform of *DISC1* (L1) and the genomic structure of mouse *DISC1*. RT-PCR between the predicted mouse exon 13 and exon 2 failed to detect any transcripts in mouse

brain or heart cDNA. However, RT-PCR within exon 2 failed to detect transcription of the experimentally verified exon 2. The same reaction did work on control genomic DNA templates demonstrating that the PCR was robust (data not shown). The failure to detect mouse *DISC1* by RT-PCR was consistent with an inability to detect transcription by Northern blot hybridisation (R. Devon, personal communication) or by *in situ* hybridisation (J. Fantes, personal communication). Previous attempts to detect mouse *DISC1* by RT-PCR were also inconsistent (data not shown and R. Devon personal communication). Combined these observations suggest that mouse *DISC1* is expressed at very low levels (at least in the C57BL/6J strain used) or in tightly restricted temporal, spatial or conditional patterns. Consequently the predicted structure of exons 9 to 13 could not be experimentally demonstrated.

Based on the known genomic structure of mouse exons 1 to 8 and the homology based prediction of exons 9 to 13, mouse *DISC1* was found to have the same intron/ exon structure as human *DISC1* (figure 6.6).

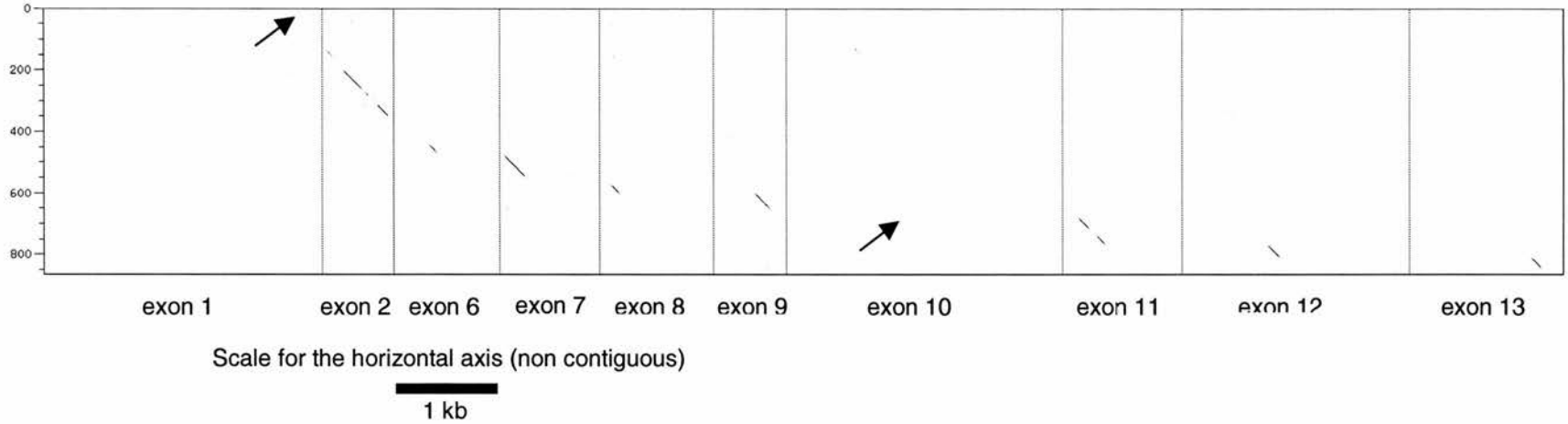


Figure 6.7; Dotmatrix alignment of human *DISC1* isoform L1 amino acid sequence to ordered and oriented sequence contigs assembled from the mouse whole genome shotgun sequence. Sequence contigs are divided by solid vertical lines. Numbers below the alignment indicate the exon number of human *DISC1* that shares sequence similarity with the associated contig. There was no sequence coverage in the whole genome shotgun sequence data set for exons 3, 4 and 5. However, sequence for mouse exons 1 to 8 had previously been generated (R. Devon). Arrows indicate the significant alignments for the short stretches of sequence representing exons 1 and 10.

6.5 Evolution of *DISC1* gene structure

6.5.1 Comparative gene structure analysis

Human and mouse *DISC1* were predicted to have the same genomic structure as each other (section 6.4.3). *Fugu* and *Tetraodon* were also predicted to have the same genomic structure as each other (section 6.4.1). The longest possible isoforms of *DISC1* in humans and mice were encoded by 13 exons. However, there was evidence for three C-terminally truncated isoforms, each with distinct 3' UTRs. In *Fugu* and *Tetraodon*, the longest possible isoform of *DISC1* was found to be encoded by 14 exons (figure 6.4).

In comparison between mammals and fish, introns 1, 2, 3, 5, 6, 7, 8, 9, 10 and 12 were all found to be in the same position relative to protein coding sequence and their splicing phase was conserved (figure 6.8). Intron 4 is a phase 2 intron in both mammals and fish but its position relative to protein coding sequence may be shifted. However, the poor conservation of encoded amino acids in this region of the gene may have resulted in the misalignment of amino acids around the intron 4 location (figure 6.8).

The principal difference in gene structure between mammalian and piscine *DISC1* genes was the additional exon (11a) between exons 11 and 12 in fish (figures 6.8 and 6.9). This additional 45 nucleotide exon was flanked by phase 0 introns so could be included or excluded from a transcript containing exons 11 and 12 without shifting the reading frame of the downstream coding sequence. This difference in genomic organisation is particularly interesting in light of the differential use of human exon 11 splice donor sites, dividing human and by inference mouse exon 11 into a proximal and distal region. The proximal region was constitutively included in the isoforms of *DISC1* that were not C-terminally truncated, whereas the distal region was alternately included or excluded (section 6.2.2). *Fugu* exon 11 aligns with the proximal region of human and mouse exon 11 but not the distal region. Exon 11a of *Fugu DISC1* shares sequence similarity (figure 6.8) with the distal region of human

exon 11. This data implied that *Fugu* exons 11 and 11a combined are orthologous to the single exon 11 of human and mouse.

An ancestral vertebrate *DISC1* gene may have had the fish exon 11 and 11a organisation, but during the lineage to mammals the intron between these exons was lost, resulting in the two separate exons becoming a single exon. Alternatively, the ancestral *DISC1* gene may have been a single exon as in mammals and an intron insertion into the exon in the lineage to Tetraodontiformes resulted in the single exon becoming two. These two possibilities can not be fully evaluated with the data set available. However, the conservation of a splice donor site in mammalian exon 11 at the site of the intron in *Fugu DISC1* suggested that intron loss in the mammalian lineage was the most plausible explanation. The conserved ability to alternatively include or exclude *Fugu* exon 11a or the orthologous distal region of mammalian exon 11 in a *DISC1* transcript suggests that this alternative splicing is functionally significant. Further consequences of this evolutionary change in gene structure are discussed in sections 6.5.2, 6.8 and chapter 10.

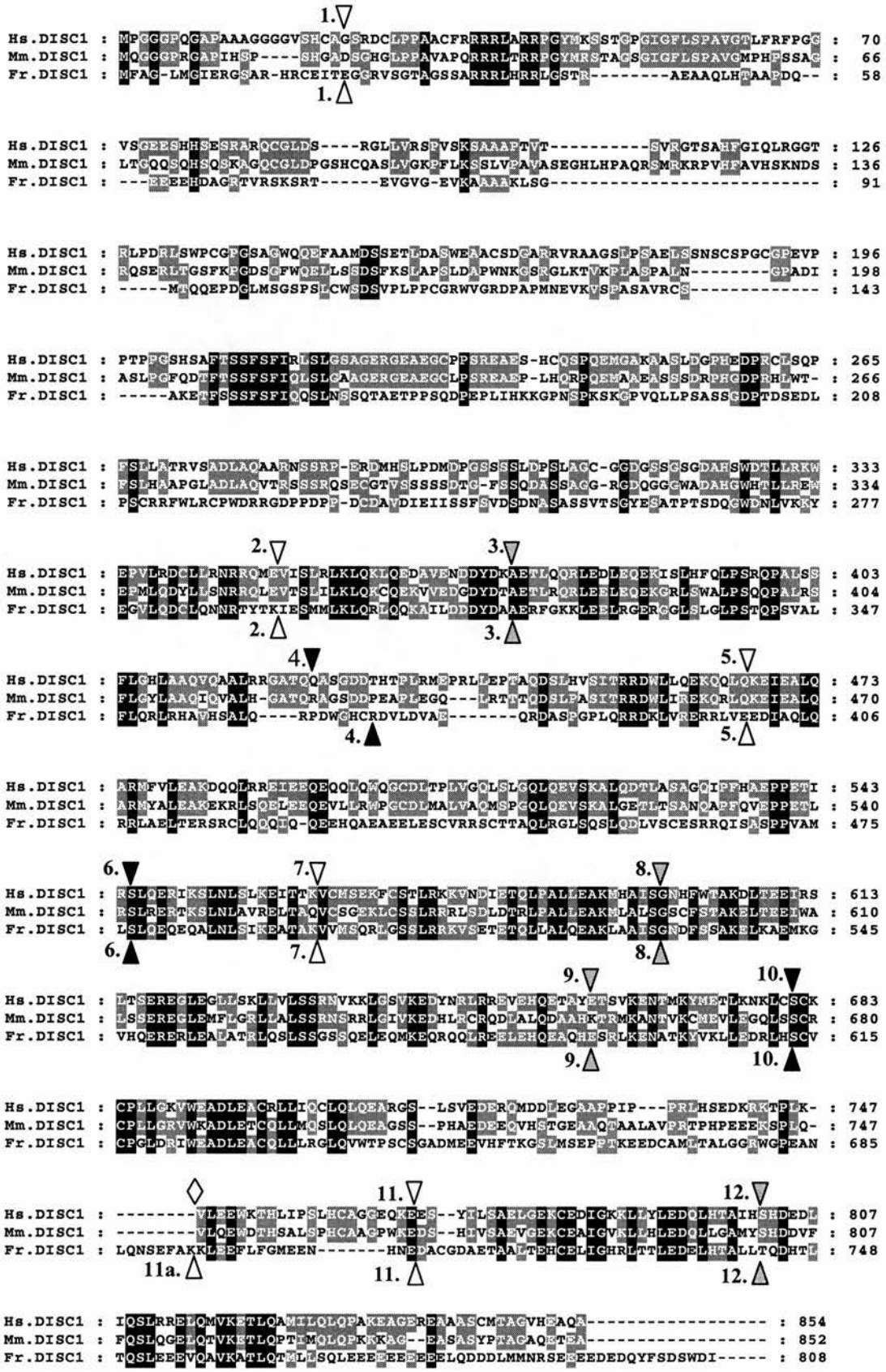


Figure 6.8; Comparative gene structure of vertebrate *DISC1*. See next page for full legend.

Figure 6.8; Comparative genomic structure of vertebrate *DISC1*. The translation of the longest known isoform of *DISC1* from human (Hs), mouse (Mm) and *Fugu* (Fr). Amino acids identical between all three species are indicated with a black background. A grey background indicates conservation between two of the three species. The alignment was created using clustalw (section 2.11.2). Where nucleic acid conservation and splice site location indicated the biologically relevant alignment was not the one reported by clustalw the alignment was manually edited. Refinements to the alignment were made around the position of intron 1 and intron 11. The position of splice sites relative to encoded amino acids is indicated by triangles. Splice site positions in human and mouse are indicated above the alignment (Human and mouse have an identical genomic structure to one another). Splice site positions in *Fugu* are indicated below the alignment. An open triangle indicates a phase 0 splice site, a grey filled triangle indicates a phase 1 splice site and a black filled triangle indicates a phase 3 splice site. The open diamond above the alignment indicates a phase 0 splice donor site in human exon 11 that is used in the L2 splice variant. Numbering adjacent to each of the splice site markers indicates the intron number as referred to in the main text, note intron 11a before intron 11 in *Fugu*. Numbers to the right of the alignment indicate the amino acid coordinate of the last residue in that block of the alignment.

6.5.2 *DISC1* splicing strategies

In humans, there are five known splice variants of *DISC1* (isoforms) that affect the protein coding sequence (section 6.2). Two of the isoforms truncate the *DISC1* open reading frame shortly after exon 9 coding sequence (S1 and S2). A further isoform truncates the reading frame shortly after exon 3 (E3 isoform).

Two short splice variants of *Fugu DISC1* that were identified by RT-PCR (section 6.3), represented splicing from the exon 9 splice donor site to exon 11a or exon 12 splice acceptor sites. Intron 9 was a phase 1 intron, whereas introns 11a and 11 were phase 0 introns. Consequently, splicing from exon 9 to exon 11a or exon 12 resulted in a frame shift of the downstream sequence. The splicing of exon 9 to exon 11a resulted in the continuation of an open reading frame for 14 codons before an in-frame stop codon. Exon 9 splicing to exon 12 resulted in the continuation of an open reading frame for only 3 codons before an in-frame stop codon. The entire length of exon 11a was predicted to encode for protein in the L1 isoform of *DISC1* and in the exon 9 to 11a splice form. However, these two isoforms would translate exon 11a in different reading frames. The third codon position of the L1 reading frame would be

the first codon position of the exon 9 to 11a splice form reading frame. It is interesting to note that both the exon 11a reading frames that were predicted to be translated were equally conserved between *Fugu* and *Tetraodon* (one non-synonymous change in each reading frame). There is no evidence to suggest that the distal region of human exon 11 is translated in more than one reading frame.

Whereas human *DISC1* was C-terminally truncated after exon 9 by splicing into an alternative terminal exon (S1 isoform) or failing to splice from the exon 9 splice donor site (S2 isoform), *Fugu* achieved a similar truncation by splicing into *DISC1* exons in combinations that altered their reading frame and introduced in frame stop codons. It is likely that the S1 and S2 isoforms of human *DISC1* are equivalent to the *Fugu DISC1*, post exon 9 truncations. There is no obvious sequence similarity between the C-terminal residues of human post exon 9 truncations and the alternate post exon 9 truncations produced by *Fugu*. *Fugu* exon 9 to 11a and exon 9 to 12 splice variants are referred to as S1 and S2 respectively. There is no inference that *Fugu* S1 is functionally equivalent to human S1 and not S2 or vice versa.

In both human and mouse, the E3 splice variant was demonstrated. This splice variant terminates the *DISC1* open reading frame two codons downstream of the exon 3 splice donor site (section 6.2). In the E3 splice variant the exon 3 splice donor site was not utilised and a 3' UTR comprising the genomic sequence directly downstream of exon 3 represented the 3' end of this transcript. No evidence was fund for such a transcript in *Fugu*. The absence of a cleavage and polyadenylation signal (AATAAA) in intron 3 further suggested that an alternative *DISC1* 3' UTR was not encoded within *Fugu* intron 3. It remains unclear as to whether *Fugu* produces an equivalent of the *DISC1* E3 isoform.

Based on the genomic structure of *Fugu DISC1*, equivalents of both the L1 and L2 splice forms could be produced by the alternative inclusion or exclusion of the 11a exon in a transcript. As exon 11a is phase neutral (flanked by introns of the same splicing phase), the downstream reading frame would be unaffected by alternative splicing of exon 11a. Exon 11a was clearly not a constitutive exon of the *Fugu DISC1*

transcript as it was found to be alternatively spliced in the S1 and S2 transcripts. However, a *Fugu* equivalent of the L2 splice form was not detected by RT-PCR in either heart or ovary RNA (the only RNA available). While this was a restricted tissue set, human L1 and L2 transcripts were readily detected by RT-PCR in heart tissue from early embryo (8 weeks) to adult and foetal brain tissue (R James and K Millar, personal communication). It is also of note that only the L1 form and not the L2 form was detected in zebrafish whole embryo by RT-PCR (section 6.4.2) although as the genomic structure had not been resolved it is not known if an L2 form could be encoded by the zebrafish *DISC1* gene(s). It may be that the L2 isoform was not produced in these species or that its expression level compared to L1 was lower in these species than in human. Another possibility is that the expression of L2 is under tighter temporal and/or spatial constraint in fish compared to humans.

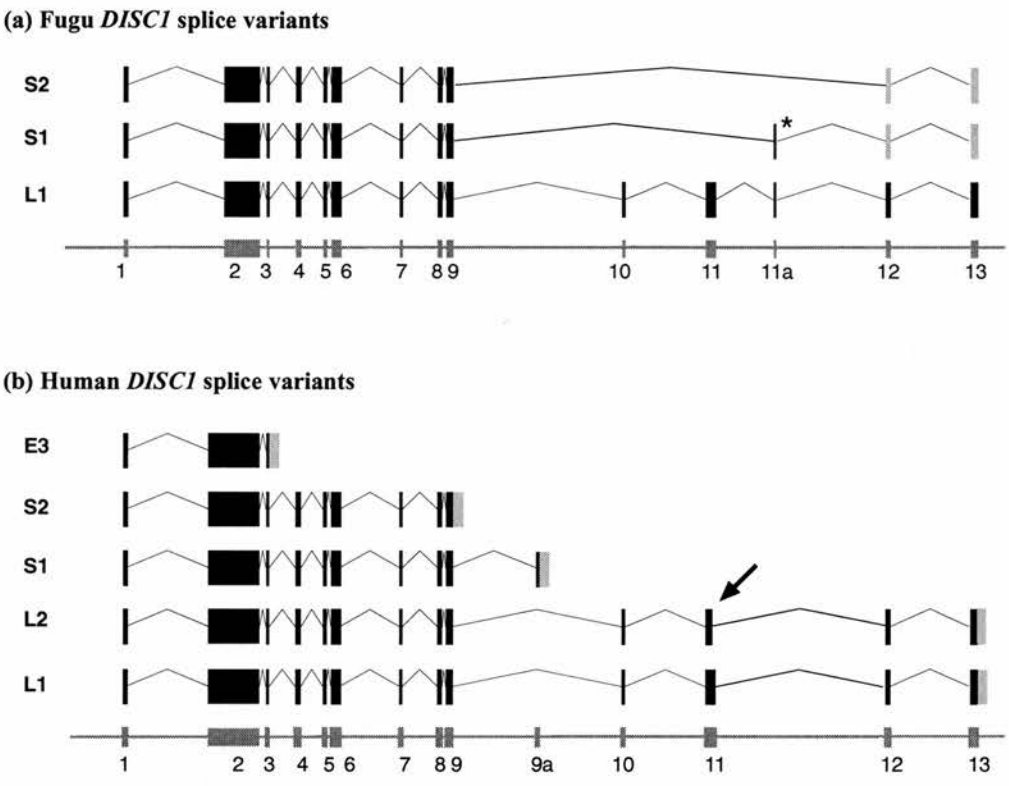


Figure 6.9; Alternate *DISC1* splicing strategies. Not shown to scale. **(a)** *Fugu* transcripts of the *DISC1* gene identified by RT-PCR (section 6.3.2). The three distinct transcripts (S1, S2 and L1) identified are shown as black and grey boxes representing protein coding and non-coding sequence respectively. An asterisk highlights exon 11a in the S1 transcript which although predicted to be protein coding, would be translated in a different reading frame to its translation in the L1 transcript. The lower grey line indicates the genomic sequence with exons indicated as numbered boxes. Introns and exons are shown to scale. **(b)** Human transcripts of the *DISC1* gene (E3, S1, S2, L1 and L2 splice forms). See section 6.2 to review the evidence for each transcript. The numbered grey boxes indicate exons in genomic sequence. Black and grey boxes connected by carat shaped lines indicate the splicing pattern of distinct transcripts. The arrow highlights the subtle difference in the size of exon 11 contributed sequence to the L1 and L2 transcripts, reflecting alternate exon 11 splice donor site usage.

6.5.3 An evolutionarily conserved AT-AC intron

The splice sites of human *DISC1* intron 12 were reported to conform to the consensus sequence of the rare AT-AC subset of introns rather than the canonical vertebrate splice site consensus sequences (Millar *et al.*, 2000a). This class of introns was first

recognised by Jackson (1991). It has subsequently been demonstrated that these introns are processed by a novel spliceosome that characteristically contains the U12 small nuclear RNA (snRNA) (Hall and Padgett, 1996; Tarn and Steitz, 1996). These "U12-type" introns account for only 0.05% of mammalian splice sites (Burset *et al.,* 2000) but are found to occur at a low frequency in a wide range of eukaryotes including vertebrates, plants and insects (Wu *et al.,* 1996).

It has been reported that through a series of one step mutational changes, a U12-type AT-AC intron can change to a canonical GT-AG intron spliced by the more common U2 spliceosome, without the intermediate loss of a functional splice site (Burge *et al.,* 1998). The key intermediate in this evolutionary change is an intron with GT-AG termini that retains the remainder of the U12 splice site sequences and is spliced by the U12 spliceosome. Such an intron is typified by the human Huntingtin (HD) gene intron 66 (Burge *et al.,* 1998). To date a U12-type AT-AC intron has not been described in *Fugu* or any other fish. In those instances where a mammalian gene contains an AT-AC U12-type intron and a fish orthologue has been described, the equivalent fish intron was found to be a canonical U2-type intron, a U12 type intron with GT-AG termini or to not contain an intron at the equivalent position within the gene (Burge *et al.,* 1998).

Alignment of *Fugu DISC1* cDNA to genomic sequence (section 6.3.2) demonstrated *Fugu DISC1* intron 12 (see figure 6.8 for *Fugu DISC1* gene structure nomenclature) is a U12-type AT-AC intron (figure 6.10). This represents the first description of such an intron in fish. Homology based gene structure prediction on mouse and *Tetraodon* sequences (section 6.4) demonstrated the intron 12 splice sites of *DISC1* are highly conserved U12-type AT-AC introns (figure 6.10).

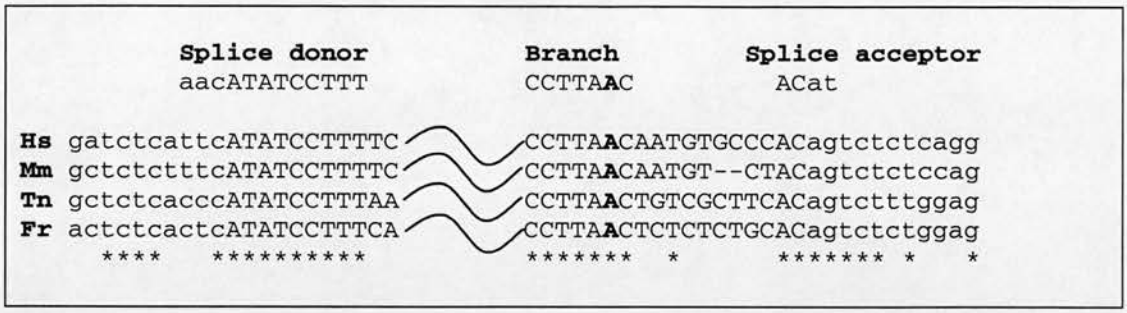


Figure 6.10; Alignment of conserved *DISC1* intron 12 splice sites. Splice donor, branch point and splice acceptor site consensus sequences for the rare U12 type introns (Burge *et al.*, 1998) are shown above the splice site alignment. Hs, indicates *Homo sapiens*; Mm, *Mus musculus*; Tn, *Tetraodon nigroveridis* and Fr, *Fugu rubripes* sequences flanking the splice sites of *DISC1* intron 12. Exonic sequence is indicated by lower case letters, intronic sequence by upper case letters. The branch adenosine is indicated in bold. Asterisks below the alignment indicate 100% nucleotide identity between each aligned sequence. The tilde shaped line between blocks of alignment indicates the remainder of intron 12 sequence from each species. The minus signs in the mouse sequence indicate gaps inserted for alignment purposes.

6.6 Comparative genomic alignment

Both local and global alignment methods under a range of parameters were used to detect biologically relevant sequence similarity between *Fugu*, *Tetraodon* and human sequences. Using the Avid global alignment method (section 2.11) with default settings failed to identify any of the known conserved exons between human and *Fugu*, but performed well in the identification of orthologous sequences between the closely related *Fugu* and *Tetraodon* sequences (figure 6.11). Using the annotation anchored method of restricting Avid alignment (section 5.5) was successful in identifying homology between every known exon with the exception of 11 and 11a.

The BLASTZ (section 2.11.2) local alignment method was used under several modes of sensitivity, search parameters and filtering (figure 6.12). The most successful local alignment method for detecting sequence similarity between known ortholgous exons between *Fugu* and human was the high sensitivity unfiltered method, although the

signal to noise ratio was high, particularly in the region of known repetitive elements (figure 5.4). The combination of a chaining model and high sensitivity searching was successful at identifying the majority of known orthologous relationships with an improved signal to noise ratio over the reporting of all hits.

6.6.1 Evolutionarily conserved sequences

In addition to the known protein coding exons, several other sites in the *DISC1* genomic region showed small blocks of sequence similarity. The splice acceptor and donor sites of exon 7 and the splice donor site of exon 12 were particularly well conserved between *Fugu*, *Tetraodon* and human (figure 6.12). Of particular interest is the 46 nucleotide conserved motif (49 nucleotides in *Tetraodon*) upstream of exon 1 in both *Fugu* and *Tetraodon* and only nine base pairs from a TSSW (section 2.11.2) predicted transcription start site (TSSW score of 4.67). In an ungapped BLASTZ alignment between *Fugu* and human genomic sequence, this motif was 72% identical (Smith Waterman score of 118). However, the aligning human sequence was from *DISC1* intron 9 rather than directly upstream of *DISC1* exon 1. It is unclear if this small block of sequence similarity represented a chance alignment or functionally relevant conservation of a sequence whose relative position has changed during the course of evolution. A greater depth of sequence from more species would help resolve such ambiguities in interpretation.

Regions within *DISC1* intron 1, intron 12 and downstream of *DISC1* were observed to be aligned between human and *Fugu* sequences and also showed high levels of conservation between *Fugu* and *Tetraodon* where there was orthologous *Tetraodon* sequence (figure 6.11). These aligning sequences were also in the same relative position in *Fugu* and human. They were not predicted as exons or represented as ESTs and gaps in the alignments were not generally multiples of three suggesting that they did not represent protein coding sequence. These conserved sequences represent candidate transcriptional regulatory sequences.

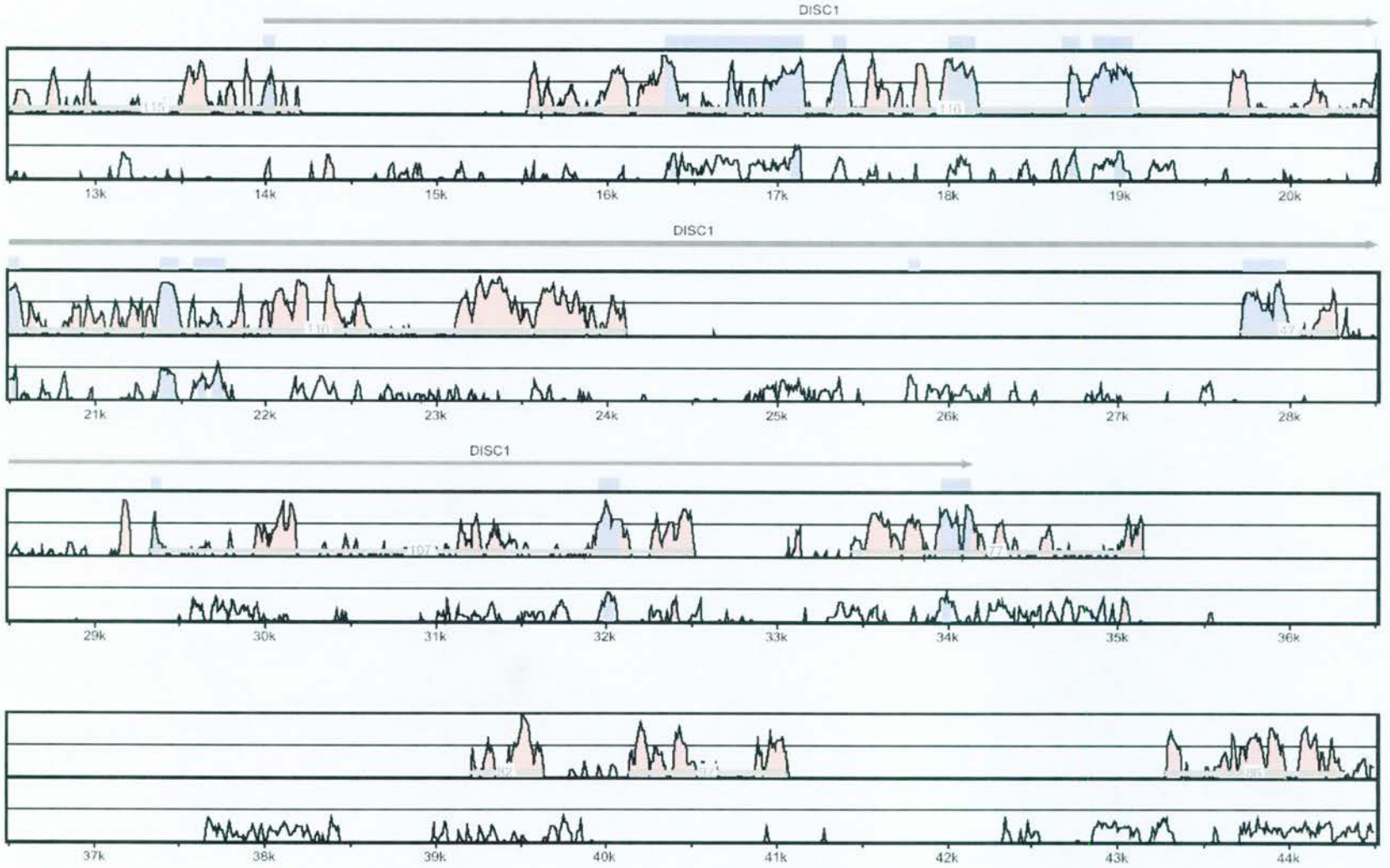


Figure 6.11; Vista plot of *Fugu* versus *Tetraodon* and *Fugu* versus human *DISC1* genomic sequence alignment. See next page for full legend.

Figure 6.11; Vista plot of genomic alignments over the *DISC1* locus. The sequence coordinates along the horizontal axis refer to coordinates of the *Fugu* genomic sequence contig (section 3.5.6). The upper plot was alignment of *Fugu* versus *Tetraodon*, the lower plot was alignment of *Fugu* versus human. Alignments were made using Avid (section 2.11.2), the *Fugu* – human alignment was 'annotation anchored' (section 5.5.1) using *DISC1* splice donor site coordinates. Vista plots were generated using a sliding window of 40 nucleotides. The minimum average identity plotted was 40%, horizontal line along the middle of each plot indicates 70% identity. Regions of >66% average identity (over 40 nucleotides) were considered candidate conserved sequences and are indicated in pink, or blue if they represent known coding sequence. Protein coding exons of the *Fugu DISC1* gene are indicated above the plots of sequence identity.

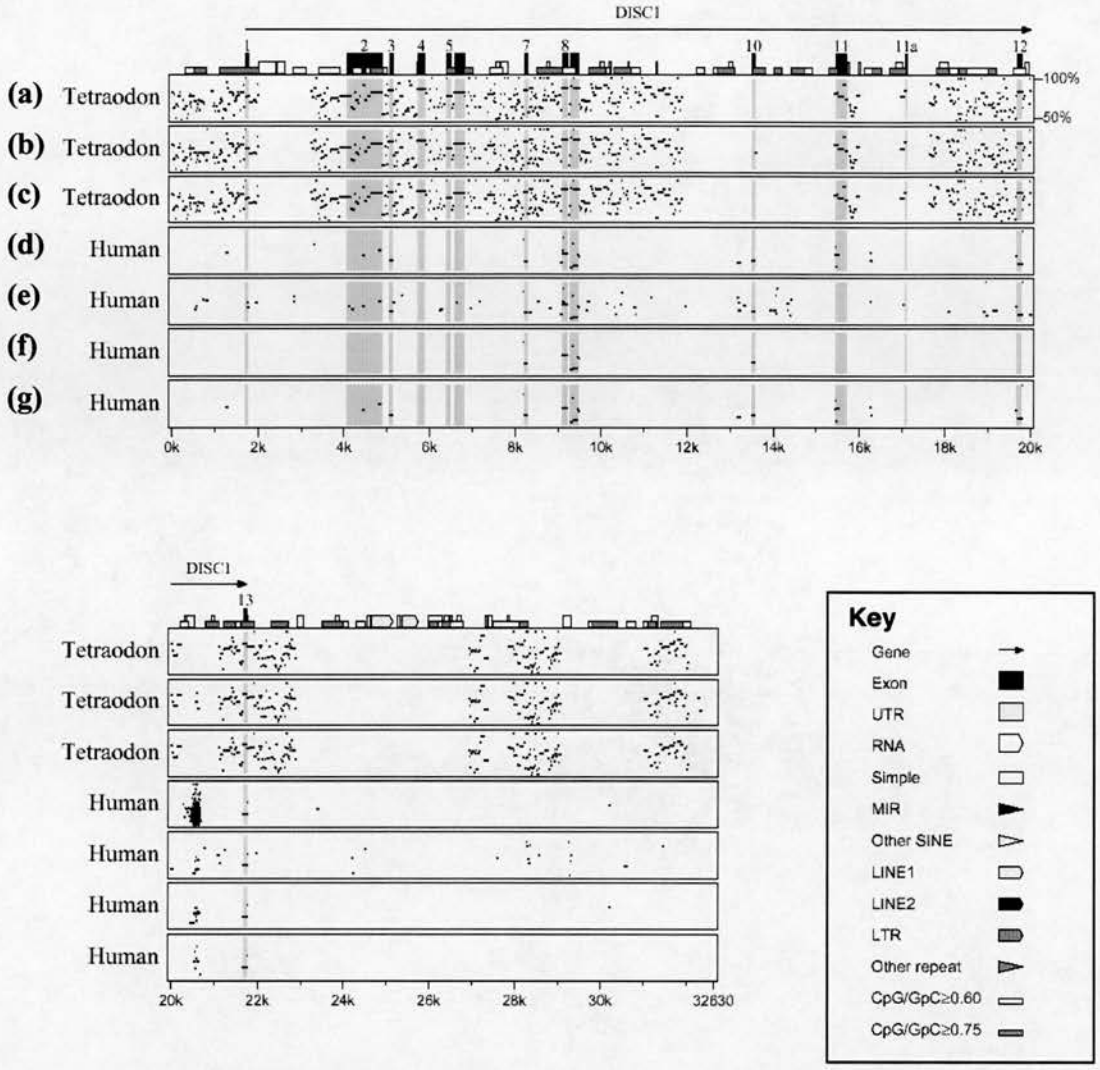


Figure 6.12; Percentage identity plot of *Fugu* versus *Tetraodon* and human *DISC1* genomic regions. The key summarises the annotation of Fugu sequence features shown above the percentage identity plots. **(a – c)** *Fugu* versus *Tetraodon* genomic sequence alignment. **(d – g)** *Fugu* versus human genomic sequence alignment. The horizontal axis coordinates relate to Fugu genomic sequence, the base sequence for each alignment. The region of *Fugu* sequence shown corresponds to the stop codon of the *TRAX* gene (section 8.4) to the end of the *Fugu* sequence contig (section 3.5.6). **(a)** All BLASTZ (section 2.11.2) matches (default settings). **(b)** BLASTZ matches with a chaining model applied to the similarity searching method. (section 2.11.2). **(c)** BLASTZ matches with a chaining model and single coverage filtering (section 2.11.2). **(d)** All BLASTZ matches (high sensitivity mode). **(e)** BLASTZ matches with a chaining model (high sensitivity mode). **(f)** BLASTZ matches with a chaining model (normal sensitivity mode). **(g)** BLASTZ matches with a chaining model and single coverage filtering (high sensitivity mode).

6.7 DISC1 protein structure

In the absence of homologous sequences for comparison, the DISC1 protein was considered to consist of a globular N-terminal region encoded by exons 1 – 2 and a helical C-terminal tail with coiled coil forming potential (section 3.2.1; Millar *et al.*, 2000a). With the depth and range of sequence provided by human, mouse and *Fugu* DISC1 sequences as well as partial sequences for zebrafish and *Tetraodon*, sequence constraints and functional predictions could be made with greater reliability and functionally conserved regions identified for further investigation. As all of the known isoforms of DISC1 reflect sub-sets of the full length isoform, comparative analysis was carried out on the full length L1 isoform and the findings related back to the truncated and alternatively spliced forms.

The original hypothesis of human DISC1 comprising two distinct regions (section 3.2.1), an N-terminal and C-terminal region, was strongly supported by comparative analysis of DISC1 amino acid sequences. The prominent compositional biases reported for human DISC1 (section 3.2.1) were found to be a conserved feature of the protein as was the distinction of N-terminal and C-terminal regions based on secondary structure prediction and levels of conservation. The N-terminal and C-terminal regions are considered separately.

6.7.1 The N-terminal head region

The N-terminal 350 amino acids of human DISC1 were poorly conserved between species. In global alignment (excluding gaps) the N-terminal region of human and mouse DISC1 were 51.5 % identical and 63.4 % similar. Human and *Fugu* were 38.6 % identical and 50.6 % similar. Conservation, particularly between Tetraodontiformes and mammals, was concentrated into 3 regions of the N-terminal domain (figure 6.14), an arginine rich motif (RRRLARRPG in human), a serine and phenylalanine rich motif (AFTSSFSFIRLSL in human) and a more loosely defined motif towards the C-

terminal end of the domain. Although the remainder of the N-terminal region showed little co-linear sequence similarity, especially between fish and mammals, there was a striking conservation of compositional bias with serine making up 13 to 15 % of residues and alanine and glycine combined making up 16 to 23 % of residues.

The extended open reading frame and probable tandem repeat expansion within exon 2 of zebrafish *DISC1* (section 6.4.2 and figure 6.13) further suggest that there are only limited sequence constraints on the N-terminal region of DISC1.

The conserved arginine rich motif conforms to a known nuclear localisation signal (Cokol *et al.,* 2000). The specific conservation of this motif in a background of poor conservation suggests that it may represent a functional nuclear localisation signal. A nuclear localisation for the known DISC1 orthologues was also predicted by the Reinhardt neural network method based on amino acid composition (Reinhardt and Hubbard, 1998).

The serine and phenylalanine rich motif defined as "F-[ST](3)-F-[ST]-F" in Prosite notation, (section 2.11.1) was only identified as a conserved protein motif in the legume lectin family of proteins (PFAM: PF00139). As the motif is within an all β-sheet protein domain of lectins and the context in DISC1 was predicted to be predominantly loop and helix structure (data not shown) it was considered unlikely that the co-occurance of this motif in DISC1 and lectins was of functional significance.

A weakly predicted transmembrane helix of human DISC1 residues 53 to 69 (TMpred score: 787) was also predicted for mouse DISC1 (residues 46 to 65), but was not conserved in *Fugu* or *Tetraodon*, suggesting that this did not represent a genuine transmembrane helix. The high serine content of this sequence resulted in many potential phosphorylation sites being predicted from the sequence (Prosite, section 2.11.1). Of the 12 *Fugu* and 13 human predicted sites, none were conserved between these species (data not shown) although the possibility that protein phosphorylation is important could not be ruled out.

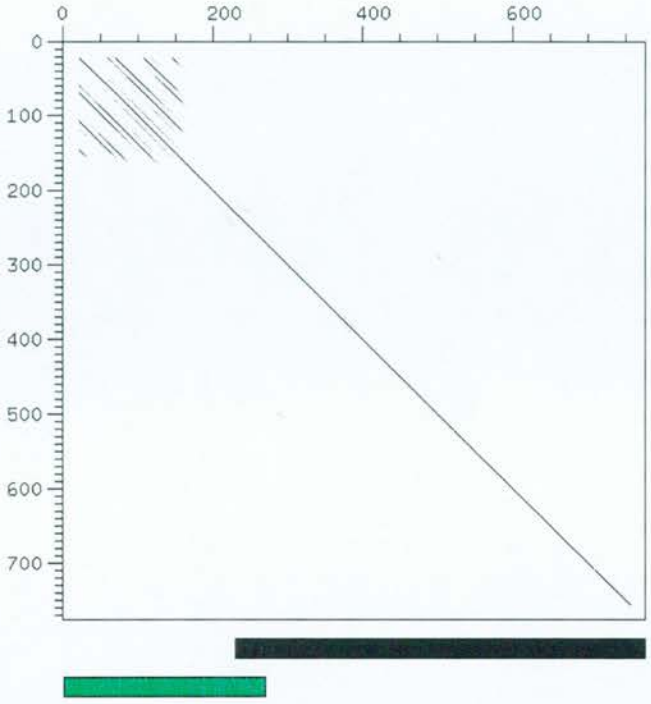


Figure 6.13; A novel amino acid repeat in the N-terminal head region of zebrafish DISC1. Dotplot of zebrafish DISC1 amino acid sequence plotted against its self with a window size of 40. The black bar below the alignment indicates the region demonstrated by RT-PCR and sequencing to represent a continuous open reading frame in a *DISC1* transcript. The green bar indicates the continuation of the *DISC1* reading frame in the assembled zebrafish genomic sequence (not experimentally verified). Within the dotplot, the block of parallel lines indicates the tandem amino acid repeat potentially encoded by zebrafish *DISC1* exon 2.

6.7.2 The C-terminal tail region

The C-terminal region of DISC1 was in general more conserved than the N-terminal region. In global alignment (excluding gaps) human and mouse C-terminal regions were 61.2 % identical and 78.1 % similar. Human and *Fugu* are 42.2 % identical and 64.9 % similar. Coiled coil forming potential was highly conserved between all orthologous sequences (figure 6.14), of particular note is the apparent modularity of coiled coil regions and the higher levels of sequence conservation within these regions than the flanking sequence (figure 6.15). There are seven consistently predicted blocks of coiled coil in the longest DISC1 isoform (figure 6.15). These blocks are subsequently referred to as blocks 'a' to 'g' with 'a' being the most N-terminal and 'g' being the most C-terminal. Blocks 'a' to 'f' were present in the isoforms that truncate after exon 9 and only block 'a' was present in the shortest isoforms that truncate after exon 3.

With the exception of a well conserved 11 amino acid motif encoded by the 5' end of exon 11, exons 11 and 12 were poorly conserved with no predicted coiled coil regions (figures 6.14 and 6.15). There was no indication at the protein level why the alternative splicing of exon 11 distal sequence / exon 11a was conserved following the putatative exon fusion event in the mammalian lineage (section 6.5.2).

Figure 6.14; Multiple sequence alignment of vertebrate DISC1 amino acid sequences. Full legend on next page.

Figure 6.14; Multiple sequence alignment of vertebrate DISC1 amino acid sequences. Hs indicates Homo sapiens; Mm, *Mus musculus*; Fr, *Fugu rubripes*; Dr, *Danio rerio*. Residues conserved between mammalian (Hs and Mm) DISC1 sequences have a brown background, those conserved between fish (Fr and Dr) have a green background. Residues conserved between all aligned sequences have a black background. Only a partial sequence was available for *Danio rerio,* the sequence shown is limited to the extent experimentally verified (section 6.4.2). Regions of coiled coil that were predicted to be present in all of the aligned sequences are indicated by a red bar above the alignment. The two conserved motifs in the N-terminal region are indicated by horizontal yellow bars above the alignment. A black triangle indicates the relative position of intron 2 that is considered the approximate boundary between N-terminal and C-terminal regions.

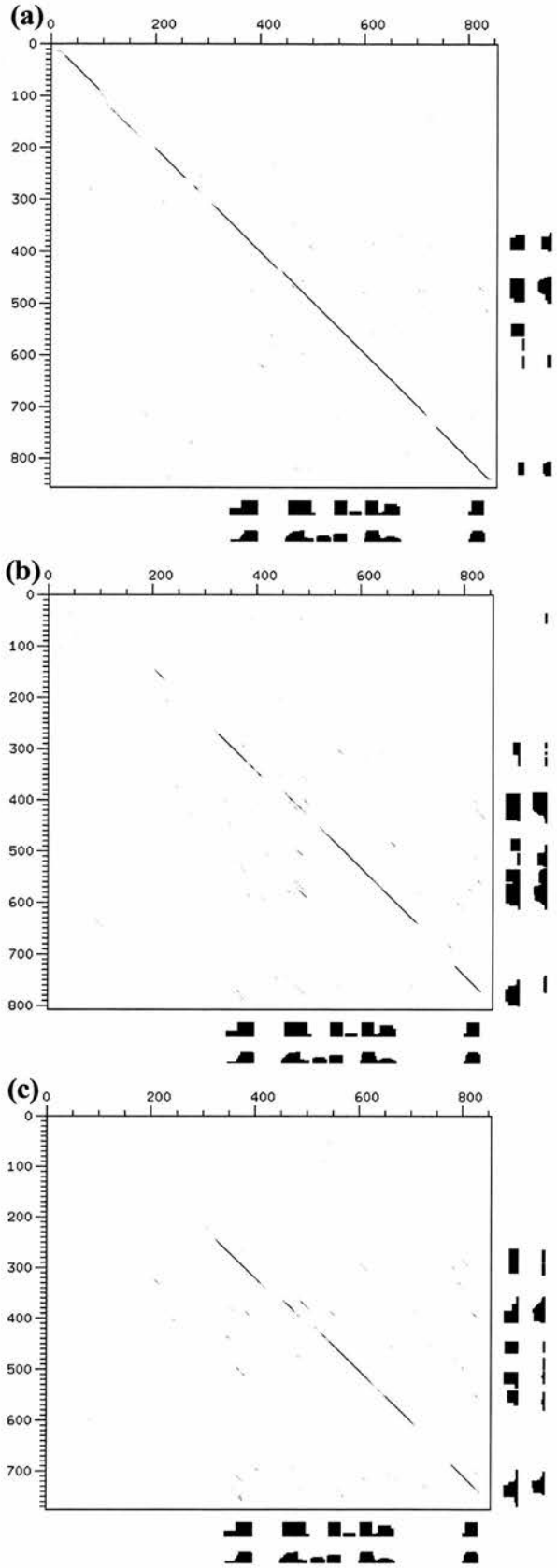


Figure 6.15; Specific conservation of predicted coiled – coil forming regions of DISC1. Dotplots (section 2.11.2) of DISC1 amino acid sequence alignments. In each case, human DISC1 is along the horizontal axis. The longest DISC1 isoform (L1) is aligned for each species. The coiled – coil prediction software, coils (section 2.11.2) and Multicoil (section 2.11.2) were used to predict coiled coil regions of each aligned amino acid sequence. Coils and Multicoils prediction of coiled – coils structures is represented as black boxes along the outside of the dotplot. Coils predictions are the inner most lane and Multicoils the outer most lane. The width of the box indicates the 'probability' score (range 0-1) for each residue of the amino acid sequence. Only scores above 0.3 are shown on the graphs. **(a)** Human versus mouse (vertical). **(b)** Human versus *Fugu* (vertical). **(c)** Human versus zebrafish DISC1. The zebrafish sequence aligned does not represent the full open reading frame (see section 6.4.2).

6.7.3 Discussion - A generalised model of DISC1

The C-terminal tail region of DISC1 was found to be the most conserved region of the protein (section 6.7.2) and is the only region of the protein known to be alternatively processed. The predicted coiled coil regions were found to be arranged into well defined and evolutionarily conserved blocks (figure 6.15).

Coiled coils are protein structures involved in protein – protein interactions that were first speculated to exist by Crick (1953) and were demonstrated to exist in mollusc tropomyosin ten years later (Cohen and Holmes, 1963). The basic structure of a two stranded coiled coil (the type predicted for the DISC1 coiled coil regions) is that of a heptad amino acid repeat where the first and fourth amino acids have hydrophobic side chains. The two alpha helical strands of amino acids then coil around one another to form a super-coiled structure with a hydrophobic core consisting of the first and fourth amino acid of the heptad repeat. The specificity of these interactions is mediated both by the particular hydrophobic residues forming the core of the coiled coil and the often polar or charged residues at the third and seventh position in the repeat (Muller *et al.*, 2000).

The modular nature of coiled coil regions in DISC1, and the alternate inclusion and exclusion of coiled coil regions in distinct isoforms of the protein (figure 6.16) suggest that DISC1 forms coiled coil interactions with a range of proteins, possibly mediating their common localisation to a multi-component complex. Preliminary findings of yeast-2-hybrid interactions with DISC1 and other coiled coil containing proteins support the prediction of multiple heteromeric interactions mediated by coiled coils (K Millar, personal communication). A logical next step in the characterisation of this protein would be the 'dissection' of each conserved coiled coil region from the DISC1 protein and its testing in a yeast-2-hybrid system. The results of such an experiment would define the protein – protein interactions that could be made by each of the DISC1 isoforms.

The specific conservation of an N-terminal nuclear localisation signal strongly implies that DISC1 is a nuclear protein or at least does localise to the nucleus under some

circumstances. The function of the remainder of the N-terminal region is unclear, the poor conservation of sequence but the maintenance of compositional bias suggest it is the general physico-chemical properties of this region that are important. Of particular note is the well conserved phenylalanine and serine rich motif in the poorly conserved N-terminal region whose function is unknown.

The ability to alternatively include or exclude distal exon 11 encoded amino acids in the long isoforms of DISC1 has been conserved in evolution even after the genomic reorganisation of the sequences involved, suggesting functional significance. At the protein level there is little indication of that functional significance. One possibility is that spacing between coiled coil blocks 'f' and 'g' has some functional importance.

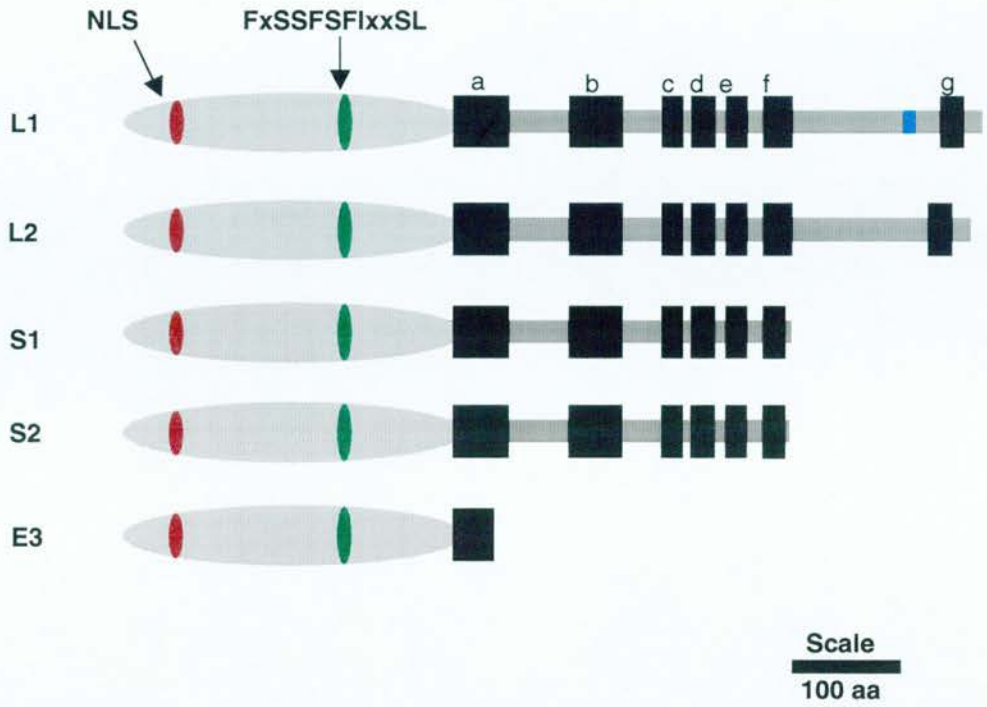


Figure 6.16; A generalised model of human DISC1 structure. The five principal isoforms of DISC1 protein that could be produced by differential RNA processing. Light grey elipses represent the poorly conserved N-terminal region encoded by exons 1 and 2. The Dark grey bar indicates the C-terminal region. Black boxes indicate the consistently predicted and conserved coiled coil regions of the protein. A blue box in the L1 isoform represents amino acids encoded by the alternatively spliced distal region of exon 11. Two conserved motifs in the N-terminal region are indicated by coloured bars, red for the motif that conforms to a nuclear localisation signal pattern, green for a phenylalanine and serine rich motif of unknown function. Letters 'a' to 'g' indicate the nomenclature for referring to the blocks of coiled coil in the main body of text.

6.8 Discussion

The work presented in this chapter has refined the transcription map around the chromosome 1 breakpoint in humans and identified three previously unknown splice variants of *DISC1*. The genomic organisation of *Fugu DISC1* was demonstrated as well as characterising three splice variants of the *Fugu DISC1* gene. The full length open reading frame of the longest *DISC1* isoforms were predicted for *Tetraodon* and mouse, and a partial cDNA sequence was obtained from a zebrafish homologue of *DISC1*.

Human and *Fugu DISC1* were found to use radically differing splicing strategies to generate what would appear on the basis of sequence analysis to be a functionally equivalent range of *DISC1* isoforms. Although a *Fugu* equivalent of the short E3 isoform was not found, and an equivalent of the L2 isoform was not detected by RT-PCR although both could in theory be produced by alternative splicing from the *Fugu DISC1* gene.

Comparative genomic alignment of the *DISC1* region identified candidate non-coding regulatory sequences conserved between human, *Fugu* and Tetraodon both in intron 12 and downstream of the *DISC1* gene. Within the sequence contigs for human, *Fugu* and *Tetraodon*, no evidence was found for a protein coding gene downstream of *DISC1*. The next known gene downstream of human DISC1 (*KIAA1389*) is at least 0.8 Mb from the end of *DISC1* (section 4.2.2). Based on the ENSEMBL 1.2 annotation of the draft human genome assembly (http://www.ensembl.org/) there are few ESTs and no predicted genes between *DISC1* and *KIAA1389*. The dearth of genes downstream of *DISC1* suggests that the non-coding conserved sequences identified in the last intron and downstream of *DISC1* are more likely to play a role in the regulation of *DISC1*, *TRAX* and *EGLN1* all of which are closer to the conserved sequences than *KIAA1389*.

The chromosome 1 breakpoint region is located in intron 8 of *DISC1*. Consequently, the derived chromosome 1 (der1) would be unable to produce the L1, L2, S1 or S2

isoforms of DISC1, although it would theoretically retain the ability to produce the E3 isoform. A second consequence of the translocation would be to move the candidate *cis* regulatory elements identified in intron 12 and downstream of *DISC1* to a location in *trans* to the remainder of *DISC1*, *TRAX* and *EGLN1*. If these conserved sequences are involved in the transcriptional regulation of *DISC1* or the upstream genes, their regulation would be disrupted.

It is assumed that the second copy of the *DISC1* locus in translocation carriers is intact and "functioning as normal". Since the phenotype of major mental illness co-segregating with the t(1;11) translocation is inherited in a partially penetrant dominant manner (section 1.6), any model of the molecular aetiology of the illness must account for the dominant mode of inheritance. A possible explanation is that of haploinsufficiency, where one intact copy of the locus is unable to produce enough gene product to allow normal functioning. An alternative explanation is that of a dominant negative mutation where the mutant gene product in some way overrides or inhibits the activity of the normal copy of the gene, examples of which include; Huntington disease (OMIM: 143100), familial fatal insomnia (OMIM: 600072) and CADASIL, an hereditary adult onset condition resulting in dementia and stroke (OMIM: 600276).

There is potential for the translocation to cause only the E3 isoform and / or aberrant isoforms truncating after exon 8 to be produced from the translocation chromosome. This scenario suggests a simple hypothetical mechanism for a dominant negative disruption of *DISC1* by the translocation. The E3 or aberrant isoforms transcribed from the translocation chromosome would sequester the proteins normally interacting with larger DISC1 isoforms, potentially undermining the function of the apparently more abundant longer isoforms (Millar *et al.,* 2000a), leading to a dominant negative effect.

Although new insight has been gained from this work into the probable sub-cellular localisation of DISC1, its generalised structure and evolutionary constraints, its exact function and the identity of interacting proteins remain to be determined. Without

greater insight into the functional roles of *DISC1* it cannot be evaluated as a functional candidate for involvement in the aetiology of major mental illness. In contrast, the disruption of *DISC1* by the translocation breakpoint and independent association with major mental illness (section 1.6) argue that *DISC1* is an excellent positional candidate for involvement in the aetiology of major mental illness.

Chapter 7

DISC2 – the antisense transcript

7.1 Preface

DISC2 was originally identified as a transcript that was directly disrupted by the chromosome 1 translocation breakpoint (section 1.6.6). The longest isoform of *DISC2* was reported to be at least 12 kb in length and there were at least three shorter variants as determined by Northern blot hybridisation (Millar *et al.,* 2000a). Northern blot hybridisation only detected expression of *DISC2* in heart tissue although transcripts could be detected by RT-PCR in human brain, kidney and spleen (Millar *et al.,* 2000a). A combination of cDNA library screening and sequencing of the identified clones, 5' RACE (rapid amplification of cDNA ends) and RT-PCR demonstrated a *DISC2* transcript to be at least 15 kb in length and antisense to exon 9 of *DISC1* (Millar *et al.,* 2000a and J. Wilson-Annan, personal communication).

7.2 Preliminary sequence analysis of *DISC2*

Alignment of the 15 kb *DISC2* transcript (EMBL:AF222981) to contiguous genomic sequence demonstrated that the *DISC2* transcript was unspliced and polyadenylated. Two consensus polyadenylation sites (AATAAA) were demonstrated by ESTs of the Unigene cluster Hs.96883 (section 2.11.1) to be utilised for cleavage and polyadenylation of *DISC2*. The orientation of poly(A) tails, polyadenylation site and the orientation of directionally cloned cDNAs relative to the aligned genomic sequence all supported the initial supposition that *DISC2* was transcribed in an anti-parallel orientation to *DISC1* (figure 6.1). Analysis of the 15 kb *DISC2* transcript sequence (known to be incomplete at the 5' end) did not identify a candidate protein coding region within the transcript. The longest open reading frame present in the transcript was 57 codons and the start codon of this reading frame was in a 'poor' context for translation initiation (Kozak, 1996). Neither was there any significant

homology (BLASTX E-value of < 0.001) to known proteins when the repeat masked *DISC2* sequence was searched against the SPTR database (section 2.11.1). Further, interspersed repetitive elements were distributed along the length of the transcript (figure 7.2), again inconsistent with this transcript or this region of the transcript being protein coding.

7.2.1 *DISC2* – a long 3' UTR?

As *DISC2* was known to be incomplete at the 5' end, it was possible that the *DISC2* transcript that had been characterised to date represented the 3' UTR of a protein coding gene and that the protein coding region was 5' of the sequence currently obtained. However, with the longest isoform at a minimum length of 15 kb, *DISC2* would appear to represent an exceptionally long 3' UTR (based on a survey of available literature).

Systematic investigations had not previously been made into the length range, average, distribution and upper limits of human 3' UTR length. As this was an important issue in the investigation of the nature of *DISC2,* a survey of available sequences was devised and carried out. A method based on SRS (section 2.11.2) to identify all human 3' UTR sequences in the EMBL database was initially developed. However, the poor quality of annotation and often incomplete extent of 3' UTR sequences was observed to artificially bias the data set to short UTR sequences often under 100 bp (figure 7.1). Although this data set was biased, excluding meaningful statistical analysis, it was useful for the identification of extreme values and hence the range of 3' UTR sizes that had been reported. The longest reported human 3' UTR was that of *Doublecortin* at 9280 nucleotides (desPortes, *et al.,* 1998). There was one human 3' UTR in the size range between 8 and 9 kb, three between 7 and 8 kb, four between 6 and 7 kb, fourteen between 5 and 6 kb and 12,174 below 5 kb in length (figure 7.1).

To over come the issues of length bias and poor quality annotation, a confidently annotated sub-set of 861 full length mRNA sequences produced by the Kazusa DNA Research Institute (http://www.kazusa.or.jp/NEDO/) were assessed. The annotated coding sequence was used to determine the coordinates and consequently length of

the 3' UTR of each transcript. Poly(A) tracts at the end of the transcripts were excluded from the calculation of 3' UTR length. For this high quality data set, the mean 3' UTR length was observed to be 2,132 (S.E.M. 1,368) nucleotides. Within this data set, the range of 3' UTR lengths was 53 to 8,105 nucleotides (figure 7.1). Therefore, if *DISC2* was a protein coding gene, it possessed a 3' UTR substantially in excess of the normal and extreme range. On this basis, it was considered unlikely that *DISC2* represented the 3' UTR of a protein coding transcript. The subsequent assembly of contiguous human sequence over the region (section 4.3.2) its primary annotation (section 5.2) and comparative genomic analysis (sections 7.7 and 6.6) have supported the conclusion that *DISC2* does not represent the 3' UTR of a protein coding gene.

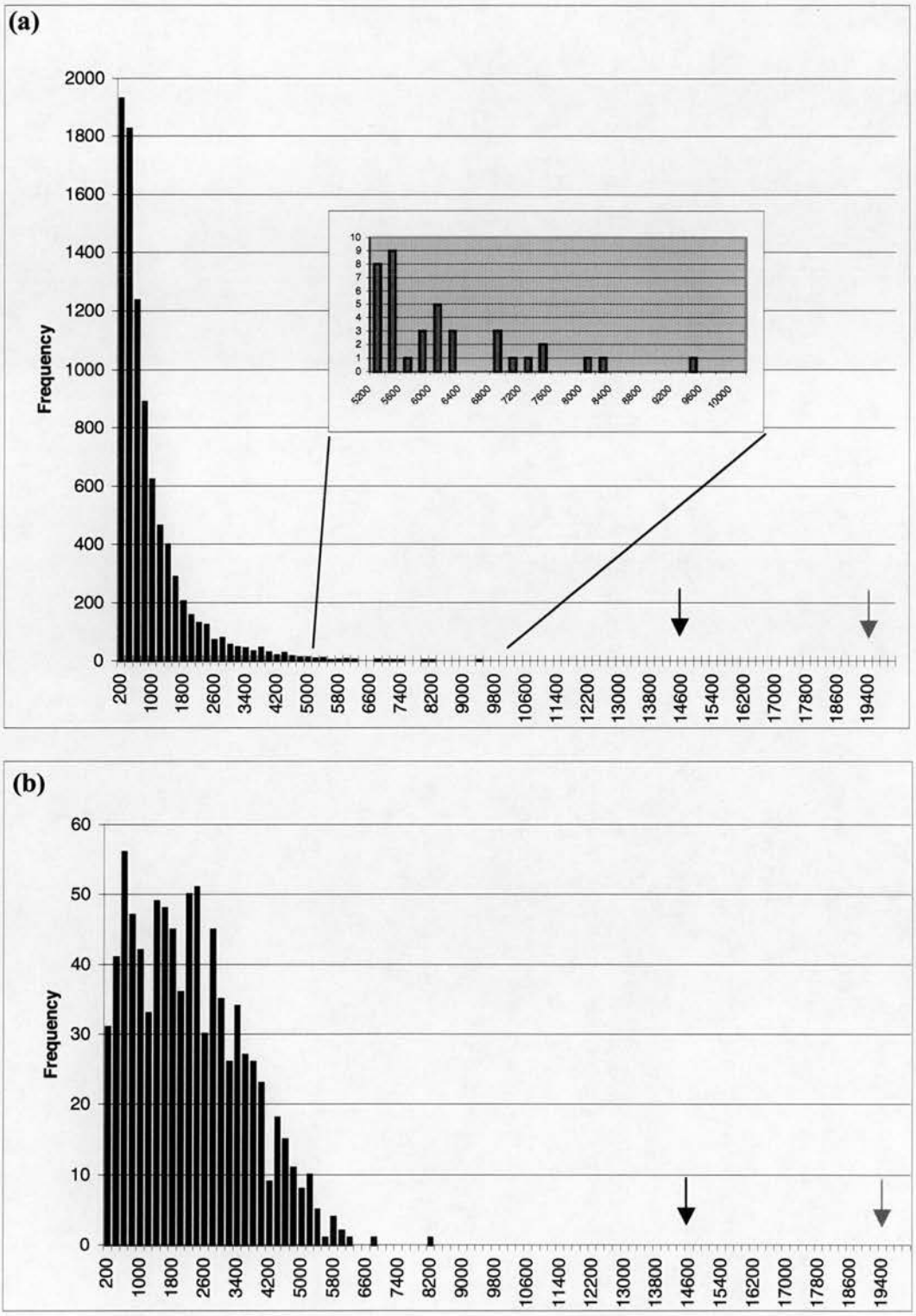


Figure 7.1;The size distribution of human 3' UTR sequences. Human 3' UTR length plotted as a frequency distribution. Each column represents a 200 bp size bin, the frequency of 3' UTR sequences with a length represented by a size bin is reflected in the height of the

column. The horizontal axis represents consecutive size bins. Numbering along the horizontal axis reflects the upper limit of the size bin. Space does not permit each size bin to be labelled. Black arrows indicate the minimum size of *DISC2* as defined previously (Millar *et al.,* 2000a). Green arrows indicate the minimum size of *DISC2* as defined in section 7.6. **(a)** The size distribution of all human 3' UTR sequences, this data set is subject to a severe acquisition bias (see section 7.2). The inset shows the small number of extreme outliers not readily visible on the main graph. **(b)** The size distribution of a confidently annotated sub-set of human 3' UTRs.

7.2.2 *DISC2* – a non-coding RNA?

If *DISC2* did not represent a protein coding transcript, by default it would then represent a new member of an emerging group of as yet poorly defined non-coding mRNA like transcripts (Erdmann *et al.,* 1999; Erdmann *et al.,* 2001 for review). The known human non-coding RNA transcripts (subsequently ncRNAs) are summarised in table 7.1. However, the total number of ncRNA transcripts is likely to be substantially higher than the small number that have been described to date. For example, in the 0.8 Mb of the human genome described in section 5.2.2 the *DISC2*, *Backtrax* and *Foretrax* transcripts (section 5.2.2) could all be considered ncRNA transcripts. They are polyadenylated and are unlikely to be protein coding (sections 5.2.2 and 8.5).

Functional roles have not generally been identified for ncRNA transcripts although several ncRNA transcripts have been reported that display spatial and temporal regulation (Leu *et al.,* 1997; Velleca *et al.,* 1994) or are induced under specific conditions such as stress (Hollander *et al.,* 1996) or viral infection (Li, *et al.,* 1997). The exceptions to this are *XIST* and *SRA*. *XIST* is a 16.5 kb (in humans) non-coding, spliced and polyadenylated RNA that is expressed only from the 'inactive' X chromosome of most adult cells of placental mammals (Avner and Heard, 2001 for review). The expression of *XIST* is necessary for the maintenance of X inactivation in XX cells, a critical gene expression – dosage compensation system. Interestingly, *XIST* itself interacts with *Tsix*, another ncRNA that is antisense to *XIST* (Mlynarczyk and Pannig, 2000 for review). The *SRA* transcript was found to be biochemically active in the regulation of steroid receptor proteins in the absence of translation,

demonstrating that this short (~800 bp) ncRNA is functionally active as an RNA molecule (Lanz *et al.,* 1999).

Other than the mRNA-like processing of ncRNAs, the only other feature shared between them is the absence of an apparent protein coding open reading frame. Of the ncRNA transcripts that have been described (table 7.1), *NTT* would appear to be the most similar to *DISC2* both in terms of length (17.5 kb versus >15 kb for *DISC2*) and its unspliced nature. Figure 7.2 summarises the identifiable sequence features of both *DISC2* and *NTT*. While both of these transcripts possessed a similar range and distribution of identifiable features, the linear organisation of features was not found to be highly similar, nor was it noticeably different from anonymous intronic sequence (personal observation). The only feature of particular note in both *DISC2* and *NTT* was the occurrence of direct repeats, an 83 bp repeat in *NTT* and a perfect 79 bp repeat in *DISC2* (figure 7.2). Neither of these direct repeats conformed to known mRNA consensus motifs (Jacobs *et al.,* 2000 for review). There was no identifiable sequence similarity between the 79 bp direct repeat of *DISC2* and the orthologous region of the *Fugu* or *Tetraodon* genomes (figure 6.6 and data not shown).

The observation of trans-splicing in *Drosophila* (Labrador *et al.,* 2001; Dorn *et al.,* 2001) raised the exciting possibility that *DISC1* and *DISC2* could be spliced in *trans* to form a single mature transcript even though they are transcribed from different strands. The *Drosophila* Mod(mdg4) locus is similar in organisation to the *DISC1* / *DISC2* locus in that a multi-exon protein coding gene is encoded on one strand and a partially antisense transcript is synthesised from the other strand. Labrador *et al.,* (2001) demonstrated that the partially antisense transcript was spliced with the transcript of the sense strand, contributing an alternative terminal exon and 3' end to the transcript. However, if this were the case at the *DISC1* locus, probes from *DISC1* and *DISC2* would be expected to hybridise to at least one common band by Northern blot analysis, the hypothetical trans-spliced product. No such common band was observed in Northern blot hybridisation (Millar *et al.,* 2000a).

Transcript	Length	S	Function	Ref.
AIR	107.8	-	Imprinting / regulation.	**a**
antiGNAS1	nd	+	Imprinting / regulation.	**b**
BC1$^{\$}$	0.2	nd	Dendritic localisation, Translin binding	**c**
CMPD	3.4	+	Testis specific, possible role in sex determination	**d**
DGCR5	7.2	+	Unknown. In DiGeorge syndrome region.	**e**
GAS5	4.1	+	snRNA host gene	**f**
H19	2.3	+	Imprinting / regulation.	**g**
HIS1	nd	nd	Induced upon viral infection.	**h**
IPW	2.2	+	Imprinting / regulation.	**i**
NTT	17.5	-	Expressed in activated CD4+ cells.	**j**
SRA	0.8	nd	Regulator of steroid receptors.	**k**
U17HG	2.1	+	snRNA host gene	**l**
U19	nd	+	snRNA host gene	**m**
XIST	16.5	+	X chromosome inactivation.	**n**

Table 7.1; Summary of human ncRNA transcripts. Length indicates the length of the mature transcript in kilobases. Column 'S' indicates if the transcript is spliced (+) or unspliced (-). The notation 'nd' indicates not determined. All of the transcripts shown are known to be polyadenylated and are assumed to be transcribed by RNA polymerase II. The '$' symbol indicates a mouse transcript that was included in the table as an introduction to discussions in chapter 8. The reference for each transcript is indicated in the column labelled 'Ref.': **(a)** Lyle *et al.*, (2000); **(b)** Hayward and Bonthron, (2000); **(c)** Muramatsu *et al.*, (1998); **(d)** Ninomiya *et al.*, (1996); **(e)** Sutherland *et al.*, (1996); **(f)** Smith and Steitz, (1998); **(g)** Brannan *et al.*, (1990); **(h)** Askew *et al.*, (1991); **(i)** Wevrick *et al.*, (1994); **(j)** Liu *et al.*, (1997); **(k)** Lanz *et al.*, (1999); **(l)** Pelczar and Filipowicz (1998); **(m)** Bortolin and Kiss., (1998); **(n)** Brown *et al.*, (1992).

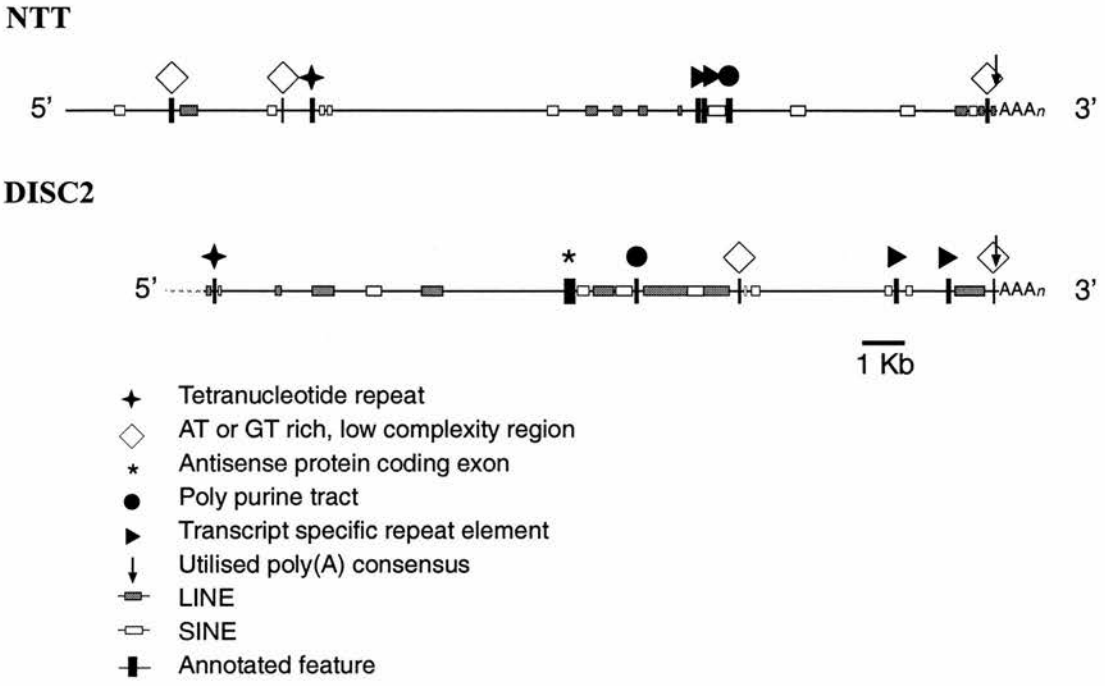


Figure 7.2; Feature based comparison of *DISC2* and *NTT* transcripts. Both transcripts represent large, unspiced, polyadenylated and apparently non-coding RNA species. The tetranucleotide repeat of *DISC2* was $(TGGA)_n$ and that of NTT was $(TGAG)_n$. Although both transcripts possess comparable features, the linear organisation of those elements is not highly similar.

7.3 Mono- / bi- allelic expression of *DISC2*

While the function of ncRNAs remains generally obscure, there is a striking correlation between genomically imprinted loci and the presence of an imprinted ncRNA (summarised in table 7.1). Genomic imprinting is the transcriptional expression of only one parental allele mediated through the epigenetic 'tagging' of gene regulatory sequences in a parent of origin dependent manner. The epigenetic tag is typically methylation of CpG di-nucleotides within an imprinting control region (ICR) (Kelley and Kuroda, 2000 for review).

Probably the best studied example of genomic imprinting is the *INS2* / *IGF2* gene cluster on human chromosome 11. The maternally derived ICR at this locus is unmethylated, this site is then bound by a chromatin insulator preventing a transcriptional enhancer from inducing the transcription of *INS2* and *IGF2*. The paternally derived ICR is heavily methylated and consequently is not bound by the chromatin insulator, allowing the expression of paternal *INS2* and *IGF2* to be induced by the enhancer. When the enhancer is unable to induce the transcription of *INS2* and *IGF2*, transcription of the ncRNA, *H19* is induced. Consequently, *H19* is only expressed from the maternally derived chromosome and *INS2* and *IGF2* only from the paternally derived chromosome (Bell and Felsenfeld, 2000; Kelley and Kuroda, 2000). Although this locus has been studied in great detail, the H19 transcript which is conserved between mammals (Onyango *et al.,* 2000) has yet to be assigned a function.

The frequent association of ncRNA transcripts and imprinted genomic loci suggested that the *DISC2* transcript could be involved in genomic imprinting. Although there was no evidence of genomic imprinting in the segregation of major mental illness through the t(1;11) family (section 1.6), it was conceivable that the translocation could disrupt imprinting which in itself could provide a mechanism leading to disease.

Allelic expression studies were designed to determine if *DISC2* was subject to genomic imprinting. Two common single nucleotide polymorphisms (SNPs), polymorphism I and polymorphism II were selected. The SNPs had previously been identified in a screen for *DISC2* polymorphisms (R. Devon, personal communication). Both polymorphisms were from the 3' end of *DISC2* and were located in non-repetitive sequence within intron 8 of the *DISC1* gene (Millar *et al.*, 2000).

An initial study was based on the extraction of genomic DNA and RNA from the leukocytes of healthy volunteers (section 2.3). PCR was carried out on the genomic DNA and RT-PCR on the RNA (with appropriate controls to ensure no genomic DNA contamination). However, *DISC2* transcription could not be detected by RT-PCR in RNA derived from leukocytes. A revised strategy used human foetal heart tissue, a tissue in which *DISC2* transcription could be detected by Northern blot hybridisation (J. Willson-Annan, personal communication; Millar *et al.*, 2000a). The tissue was obtained from Leslie Wong, MRC tissue bank. Genomic DNA and RNA were extracted from the heart tissue and PCR amplified as previously described. Both genomic and reverse transcription PCR was robust on these nucleic acid templates (figure 7.3).

For each of five human foetal heart samples (sex unknown), RNA and genomic DNA was extracted in parallel. The DNA contamination was removed from the RNA prior to its reverse transcription. Genomic and reverse transcription PCR over polymorphism I and polymorphism II was carried out (figure 7.3). Each PCR product was then sequenced from both ends so that both sequences would be good quality over the site of the common polymorphism. For polymorphism I, three genomic heterozygotes were identified, all three were found to express both parental alleles (figure 7.4; table 7.2). For polymorphism II, two genomic heterozygotes were identified. Both were found to express both parental alleles as summarised in table 7.2. These results demonstrated that *DISC2* was bi-allelicly expressed in foetal heart tissue. Therefore *DISC2* is not genomically imprinted at least in foetal heart

tissue, the principal known site for *DISC2* transcription (figure 7.7; Millar *et al.*, 2000a).

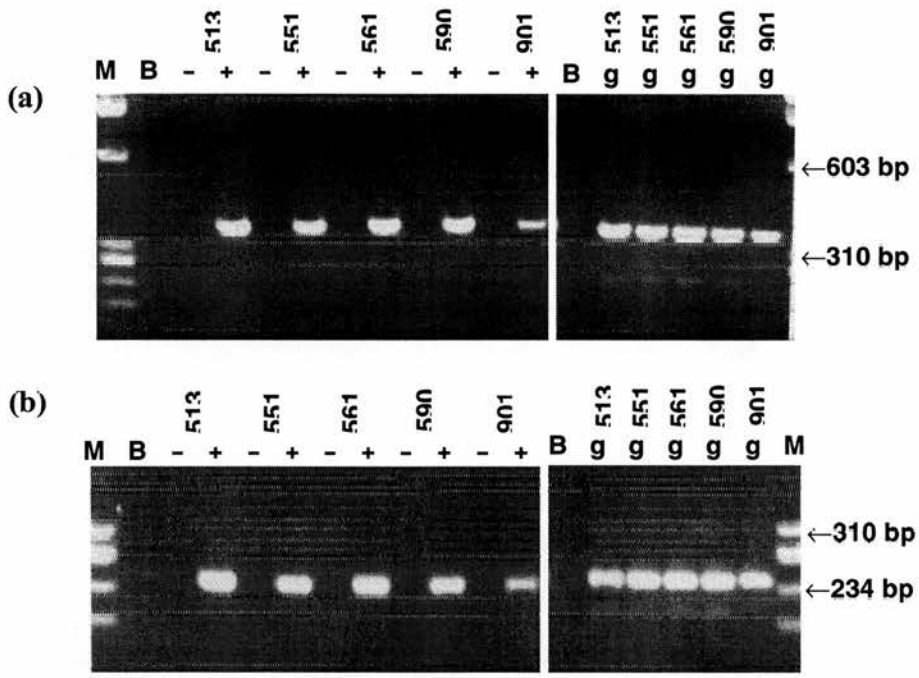
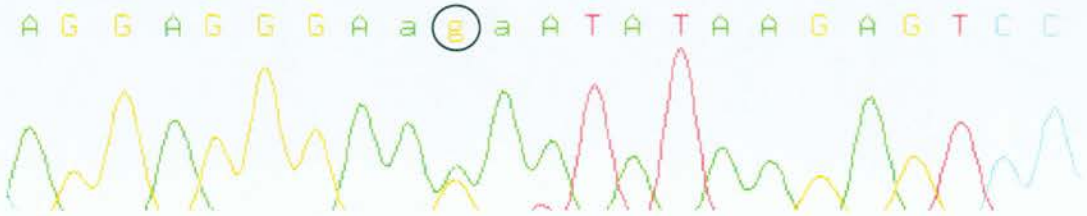


Figure 7.3; RT-PCR and genomic PCR amplification of *DISC2* sequences over the site of common polymorphisms. Sample numbers (3 digits) indicate the source tissue of RNA and genomic DNA. For RT-PCR reactions, '-' indicates a reverse transcription negative control for the RNA sample, a '+' indicates the reverse transcribed template was used for PCR. Reactions using genomic DNA as a template are marked with 'g'. Lanes marked 'B' are template negative PCR control reactions. Lanes marked 'M' indicate the ϕX174 *HaeIII* DNA molecular weight maker, relevant marker sizes are indicated to the right of the image. **(a)** PCR amplification of 350 bp of sequence over a common G/A single nucleotide polymorphism (polymorphism I). **(b)** PCR amplification of 243 bp of sequence over a common G/A single nucleotide polymorphism (polymorphism II).

(a) Sample 513 genomic PCR template

(b) Sample 513 cDNA PCR template



(c) Sample 561 genomic PCR template

(d) Sample 561 cDNA PCR template

Figure 7.4; Bi-allelic expression of *DISC2*. Genomic DNA and cDNA derived from the same foetal heart tissue sample were PCR amplified and sequenced. The position of a common polymorphism (polymorphism I) is indicated by a circle. **(a)** Chromatogram of sample 513 genomic DNA PCR product sequence. The double chromatogram peak of yellow (G) and green (A) at the position of the polymorphism indicates that sample 513 is a heterozygote for polymorphism 1. **(b)** Chromatogram of sample 513 RT-PCR product sequence. The double chromatogram peak of yellow (G) and green (A) at the position of the polymorphism indicate that both genomic alleles are transcribed. **(c)** Chromatogram of sample 561 genomic DNA PCR product. The single yellow (G) peak at the site of the polymorphism demonstrates that sample 561 is homozygous at this polymorphism, and consequently uninformative in determining differences in allelic expression. **(d)** Chromatogram of sample 561 RT-PCR product sequence.

Sample	Polymorphism I		Polymorphism II	
	Genotype	Expressed alleles	Genotype	Expressed alleles
513	GA	GA	GA	GA
551	GA	GA	G	G
561	G	G	G	G
590	GA	GA	G	G
901	A	A	GA	GA

Table 7.2; Summary of mono-allelic / bi-allelic expression studies on human foetal heart tissues. Genotypes and expressed alleles were determined by sequencing (section 2.6.5) of PCR and RT-PCR products. Each product was sequenced in both directions, for every sample the forward and reverse sequence was found to correlate. Allele scoring was based on manual inspection of sequence chromatographs. For each sample, genomic DNA and total RNA were isolated in parallel as described in section 2.3.4).

7.4 The HERV hypothesis

During primary annotation of the assembled human genomic sequence (section 5.2.3) it was noted that intron 9 of *DISC1*, which contains the expected transcriptional start site of *DISC2*, contained an integrated retrovirus of the HERV-H family. HERV-H retroviruses are sub-divided into three families (I, II and Ia) based on the organisation of repeats in the long terminal repeats (LTRs) of the provirus. The intron 9 HERV was of the sub-type II.

Within the LTR of HERV-H retroviruses, there is a core promoter consisting of a TATA box, Sp1 binding sites, a GC/AC box and Myd transcription factor binding sites. These promoters are typically uni-directional and as both LTRs of a retrovirus are direct repeats, both initiate transcription in the same direction. The orientation of the HERV-H endogenous retrovirus in *DISC1* intron 9 is such that both promoters are in an appropriate orientation to direct the transcription of *DISC2*. Several instances of retroviral promoters being utilised to direct the transcription of human genes have previously been reported. These included *ZNF80* (DiCristofano *et al.,* 1995), PLA2L (Feuchter-Murthy *et al.,* 1993), *HHLA1* and *OC90* (Kowalski *et al.,* 1999). Both *HHLA1* and *OC90* were found to be expressed from the same retroviral LTR promoter. The use of a retroviral LTR promoter to direct the transcription of

HHLA1 and *OC90* is particularly interesting as the LTR promoter in this case is of the HERV-H family (Kowalski *et al.,* 1999). *In vitro* assay has also demonstrated that HERV-H retroviral LTR promoters can direct transcription without the need for enhancers or other *cis* regulatory elements (Nelson *et al.,* 1996). The identified promoter elements of HERV-H LTR sequences (Nelson *et al.,* 1996) were found to be intact for both of the LTR sequences of the DISC1 intron 9 HERV.

The HERV-H sub-type II repeat family underwent primary expansion in the primate lineage approximately 35 million years ago. However, the high degree of sequence similarity (~92% identity) between the intron 9 HERV and at least 4 distinct HERV-H proviruses in the draft human genome sequence suggest that there have been relatively recent retrotranspositions of this retroviral sub-family in the primate lineage. Under the assumption of no selection, it has been proposed that LTR retroviral integration times can be estimated by measuring the divergence of the two LTR sequences (Lebedev, 2000). On this basis and assuming a mutation rate of 0.13% per million years (Lebedev, 2000; Medstrand and Mager, 1998) the intron 9 HERV element was predicted to have integrated to its present location approximately 17 million years ago (4.4% divergence between 5' and 3 LTRs, 4.4 / (0.13 × 2)), around the time of human and gorilla lineage divergence. Consequently, if the HERV element is involved in the transcriptional regulation of *DISC2* it is likely to represent a very recent evolutionarily event and therefore highly species restricted.

The HERV element of *DISC1* intron 9 was considered to represent a good candidate promoter for the *DISC2* transcript. The central hypothesis was that the core promoter activity of the LTR could provide the basal level of transcription detected in most cell types by RT-PCR (section 7.4.1) and that the higher levels of transcription observed by Northern blot in the heart (Millar *et al.,* 2000a; figure 7.7) reflect the effect of an unidentified heart specific enhancer.

7.4.1 Testing the HERV hypothesis

To test this hypothesis primers for RT-PCR were designed from the assembled genomic sequence contig (section 4.3) to produce a tiling path of PCR products

from the 5' most sequence known to be in the *DISC2* transcript to the first encountered LTR of the HERV element (figure 7.5). The RT-PCR reactions were carried out on human foetal heart cDNA. Every RT-PCR reaction was accompanied by a negative reverse transcription control to ensure against genomic contamination resulting in false positive results (figure 7.6). Each of the 15 RT-PCR reactions gave a positive result (figure 7.6), indicating that the HERV element could be the promoter of the *DISC2* transcript. This result also extended the minimum length for the longest isoform of *DISC2* from 15 kb to 19.7 kb. It is important to note that the Northern blot hybridisation results obtained for *DISC2* (Millar *et al.,* 2000a) could not resolve products of greater than 12 kb. Consequently an upper limit on the expected size of the longest *DISC2* isoform could not be estimated.

Additional control RT-PCR reactions were carried out beyond the HERV element (figures 7.5 and 7.6). These also produced products of a size consistent with an unspliced transcript. Reverse transcription negative control reactions did not produce products (figure 7.6) demonstrating that the PCR products did not represent contamination of reactions by genomic DNA. A sub-set of the reactions shown using foetal heart RNA (figure 7.6) were also carried out on RNA samples from other tissues. Reactions using foetal and adult brain, poly(A) selected heart, foetal limb, liver and kidney also produced products of the expected size (data not shown), demonstrating that these transcripts were not restricted to foetal heart tissue.

The results presented in this section indicate that the *DISC1* intron 9 HERV-H element may not be the promoter of the *DISC2* transcript. If the *DISC2* transcript is unspliced along its entire length and the RT-PCR product 'D' (figures 7.5 and 7.6) genuinely represents the *DISC2* transcript, the minimum size of the *DISC2* transcript would be in excess of 140 kb. This strategy of RT-PCR walking has been used previously to define the extent of partially antisense ncRNA transcripts (Lyle *et al.,* 2000). However, there is an underlying concern that the transcript being detected by RT-PCR is the unspliced *DISC1* transcript rather than the antisense *DISC2* transcript.

Possible ways to approach this problem in future work involve the RNase protection assay and strand specific probe hybridisation onto Northern blots. Single stranded probe hybridisation of Northern blots is a particularly attractive approach as the strand and approximate size of the transcript would be resolved. The use of multiple probes could then be used to map the 5' end of the transcript. Aligning the hybridisation pattern of multiple probes could then be used to infer whether the same transcript was hybridised by distinct probes. Preliminary work has demonstrated that a single stranded riboprobe can be used to detect the *DISC2* transcript and has provided the first direct evidence that *DISC2* is transcribed from the opposite strand to *DISC1* (figure 7.7). However, the largest *DISC2* transcript of more than 12 kb (Millar *et al.,* 2000a) was too faint to detect. Future work could develop this methodology into a more robust approach for defining the 5' extent of *DISC2*.

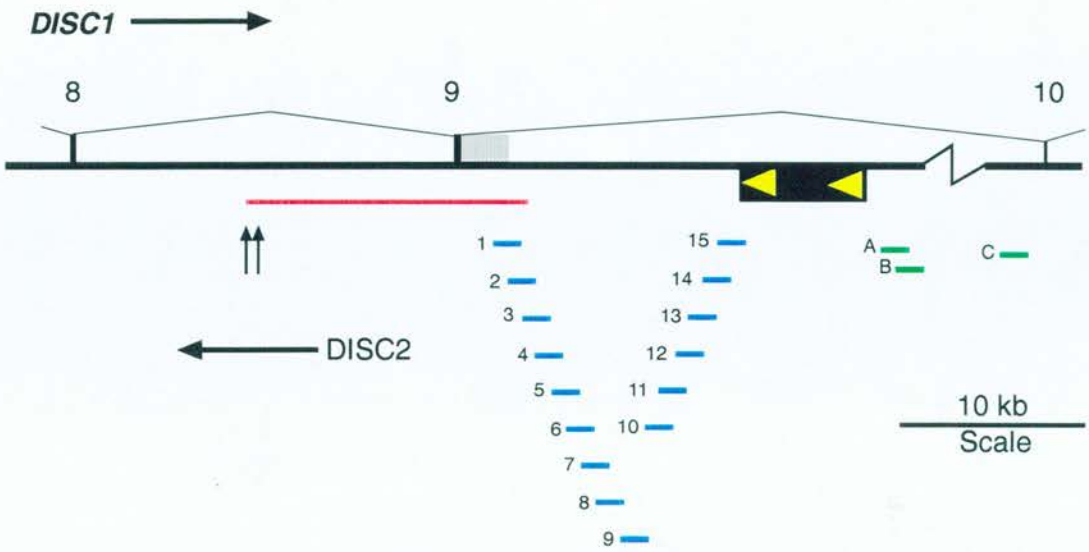


Figure 7.5; Mapping the 5' end of *DISC2*. Numbered black boxes indicate the exons 8 to 10 of *DISC1*. Carat shaped lines between exons indicate splicing of the *DISC1* transcript. The grey box to the right of exon 9 indicates an alternate 3' UTR of *DISC1* that overlaps the *DISC2* transcript. The horizontal red line shows the extent of the *DISC2* transcript that had been previously reported (Millar *et al.*, 2000a). Blue lines indicate the extent of RT-PCR reactions shown in figure 7.6. Reactions are numbered 1 to 15, the numbers correspond to the gel photos shown in figure 7.6. Horizontal green lines also indicate RT-PCR reactions 'A', 'B' and 'C' (left to right respectively). All of the RT-PCR reactions indicated were positive on human foetal heart cDNA (figure 7.6). The black filled box with two yellow triangles indicates the HERV-H element with the direct LTR repeats oriented such that they may represent the transcriptional origin of *DISC2*. The transcriptional orientation of *DISC1* and *DISC2* is indicated by labelled arrows. Vertical arrows indicate the two cleavage and polyadenylation signals utilised by the DISC2 transcript. The zig-zag line indicates that *DISC1* intron 9 is not shown to scale (the intron is 140 kb rather than the 3 kb indicated by the scale of the graphic).

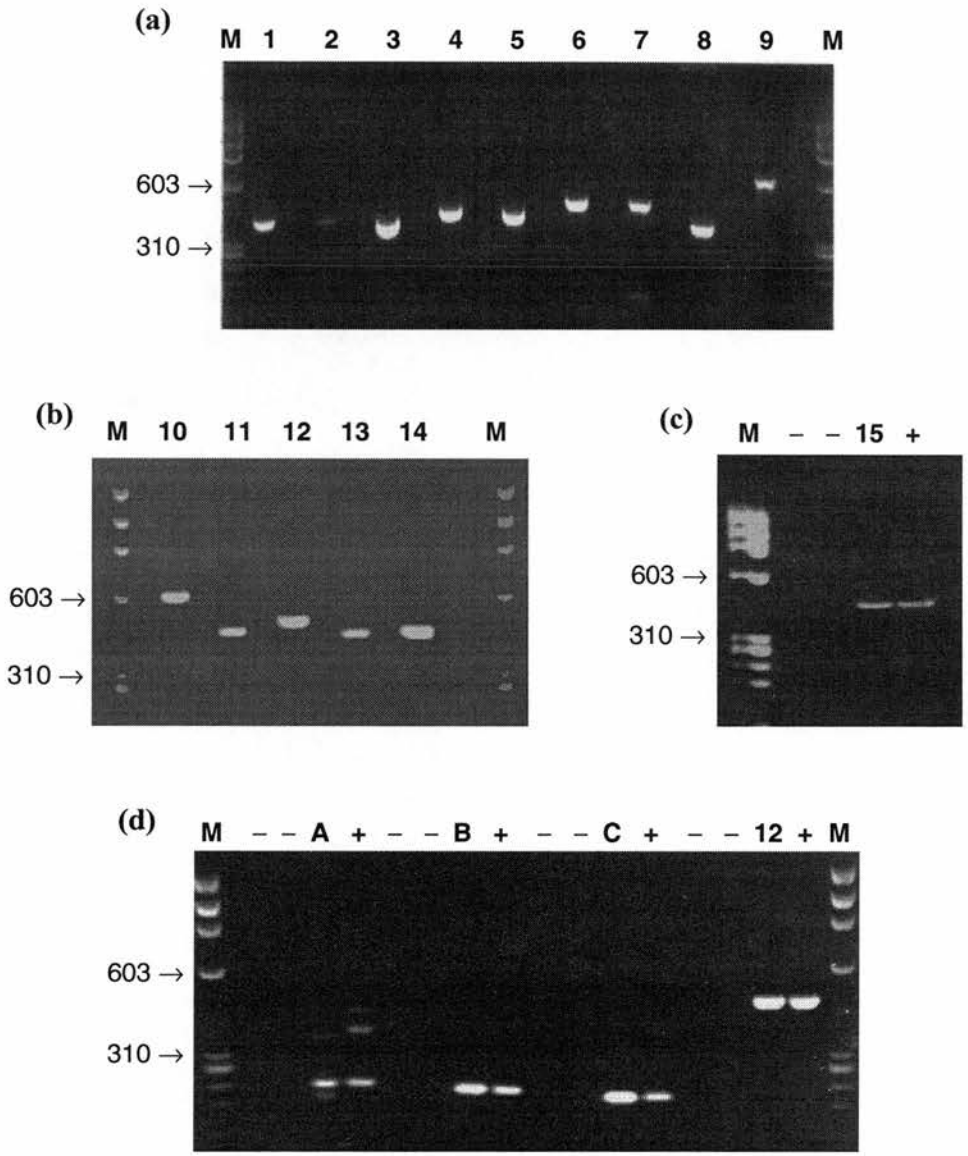


Figure 7.6; *DISC2* 5' mapping RT-PCR reactions. For gel images 'a' to 'd', 'M' indicates the DNA molecular weight marker ϕX174 was loaded on the gel, the 603 and 310 bp bands of this marker are indicated on each gel. **(a)** Results of RT-PCR reactions (1 to 9) indicated in figure 7.5. The marked lanes represent reactions using reverse transcribed foetal heart RNA as a template (RT positive). Lanes to the right of those marked indicate the foetal heart RNA had not been reverse transcribed (RT negative). Although not as abundant as other PCR products, there is a faintly visible band in the lane marked '2'. All of the PCR products are approximately the size predicted from genomic DNA sequence. **(b)** Results of RT-PCR reactions (10 to 14) indicated on figure 7.5. RT positive lanes (foetal heart RNA)

are indicated by the reaction number (10 to 14). Lanes to the left of the RT positive lanes contain the respective RT negative reactions. (c) Reaction 15 that overlaps the LTR of the HERV element. The lane marked '15' indicates the RT postitive reactions, the lane marked '+' indicates a genomic DNA template positive control reaction. Lanes marked '–' indicate the PCR negative and RT negative control reactions. (d) RT-PCR reactions A, B and C (figure 7.5). Lanes marked 'A', 'B' and 'C' represent RT positive reactions. Lanes marked '+' indicate genomic PCR controls of the RT reactions. Lanes marked '–' indicate PCR and RT negative controls. Reaction 12 from panel 'B' was also repeated at the same time using the same RNA sample demonstrating co-expression of the overlapping sequences in panels 'A' to 'C' with the non-overlapping reactions of panel 'D'.

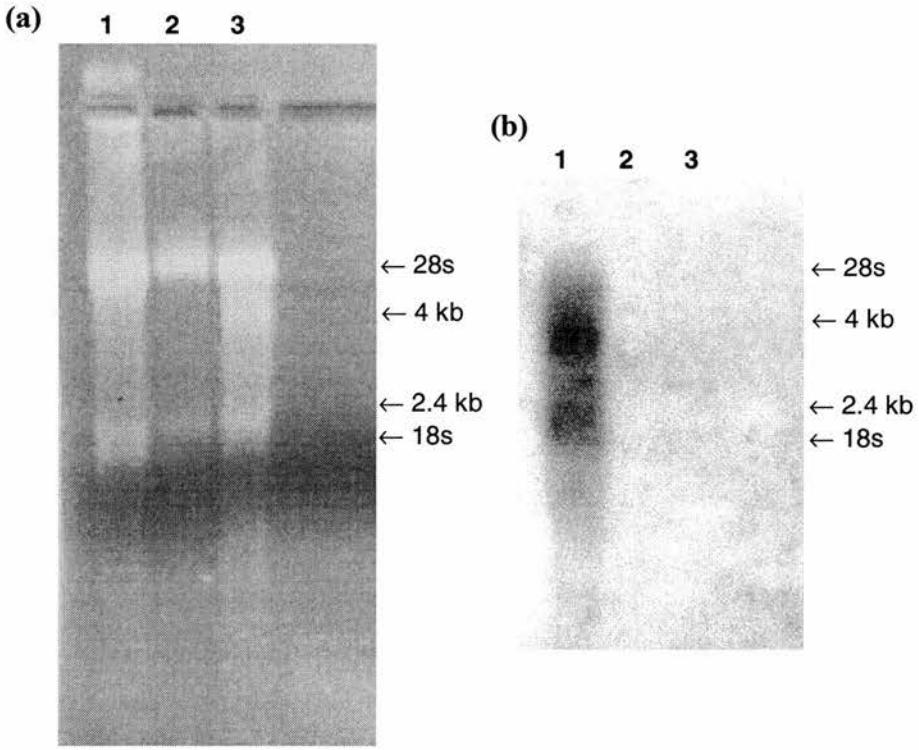


Figure 7.7; Single stranded riboprobe detection of the *DISC2* transcript. **(a)** Total cellular RNA from human foetal heart and the human fibroblast cell line 1HD (Bridger *et al.*, 1998) separated by denaturing agarose gel electrophoresis. '28s' and '18s' indicate the abundant ribosomal RNA species on the gel. Lane '1' is total human foetal heart RNA; lane '2' total human fibroblast RNA; lane 3 is total mouse foetal heart RNA. **(b)** Hybridisation of a Northern blot (of the gel shown in panel 'a') with a single stranded riboprobe. The probe was anitisense to the predicted *DISC2* transcript. A probe of complementary sequence was hybridised in parallel to a Northern blot prepared in parallel to the one shown, no signal was detected on this parallel control experiment.

7.5 *DISC2* comparative genomics

The majority of the *DISC2* transcript that is currently defined (section 7.2 and 7.6) is transcribed from the 18 kb intron 8 of human *DISC1*, although its transcriptional origin lies in intron 9 or further telomeric to the chromosome 1 breakpoint (section 7.6). Within the *Fugu* genomic sequence contig, *DISC1* intron 8 was only 86 nucleotides in length (209 fold compaction in *Fugu* relative to human). This substantial reduction in the intron size significantly reduced the complexity of searching for conserved sequence elements. *DISC1* intron 8 was not well conserved between *Fugu* and *Tetraodon* (61% identity compared to 91% for the flanking exons) and there was no identifiable sequence similarity between human and *Fugu* (section 6.6). Of particular note was the absence of consensus cleavage and polyadenylation signals within the *Fugu* or *Tetraodon* intron 8 sequence.

The only other identified feature of *DISC2* that could be specifically searched for was the 79 nucleotide direct repeat sequence towards the 3' end of *DISC2* (figure 7.2). No sequence similarity to the *DISC2* direct repeats was identified. Oligonucleotide hybridisation of the 79 bp repeat onto a mouse genomic DNA Southern blot and mouse gridded PAC clone library did not detect conservation of the repeat in the mouse genome (data not shown).

The existence of a *DISC2* transcript antisense to *Fugu DISC1* was directly tested for by RT-PCR. *Fugu* heart RNA was amplified using primers flanking *DISC1* exon 9. While *DISC1* transcripts could be detected by RT-PCR and the exon 9 flanking reaction worked on a *Fugu* genomic DNA template, no product was obtained from the equivalent RT-PCR reaction (data not shown). This result suggests that there is not a *DISC2* equivalent transcript in *Fugu*, although only one tissue was tested at one developmental stage (adult).

7.6 Discussion

DISC2 appears to be an non-protein coding mRNA-like transcript (ncRNA) that is partially antisense to the DISC1 protein coding gene (section 7.4). The *DISC2* transcript was of particular interest as it is directly disrupted by the chromosome 1

translocation breakpoint. However, as for the majority of ncRNA transcripts (table 7.1) the function, if any, of the gene product remains elusive. The common finding of such transcripts associated with genomic imprinting was interesting and lead to the question of whether the *DISC2* transcript itself was imprinted. Allelic expression studies demonstrated that both parental alleles were expressed in heart tissue, the primary known site of *DISC2* expression.

An integrated endogenous retrovirus (HERV) within intron 9 of *DISC1* was identified by analysis of assembled human genomic sequence. The HERV element was located directly upstream of the known *DISC2* transcript with two intact viral promoters oriented appropriately to direct the transcription of *DISC2*. This finding lead to the hypothesis that the HERV provided core promoter activity for the *DISC2* transcript. By RT-PCR walking, transcription was demonstrated up to the HERV element. However, multiple positive RT-PCR results up to 100 kb beyond the HERV suggested that the HERV may not provide the promoter activity of *DISC2*. The analysis of these results was complicated by the anti-parallel transcription of *DISC2* relative to *DISC1*. While control reactions unambiguously demonstrated that genomic contamination was not present, there was no control reaction to ensure that the unspliced *DISC1* hnRNA could not be detected by RT-PCR. The obvious control would be a cell type that expressed *DISC2,* but not *DISC1*. However, no such cell type is known.

The nature and function of *DISC2* remains to be fully determined. Preliminary findings have suggested that the hybridisation of single stranded riboprobes to Northern blots of heart RNA could be developed into a method for investigating the extent of the *DISC2* transcript. This approach would also allow investigation into the shorter and more abundant isoforms of the *DISC2* transcript (Millar *et al.,* 2000a). Mouse sequence that is currently being generated in the *TRAX – DISC1* orthologous region (section 4.4.4) may also provide a means for the further investigation of the *DISC2* transcript.

Chapter 8

Sequence analysis of the *TRAX* gene

8.1 Preface

The Translin Associated factor X (*TRAX*) gene is located directly upstream of human *DISC1* (section 4.2) and is intergenically spliced with *DISC1* (Millar *et al.*, 2001, see section 5.2.2). The proximity of *TRAX* to the t(1;11) translocation breakpoint and intergenic splicing of *TRAX* with *DISC1*, a transcript directly disrupted by the breakpoint necessitate the consideration of *TRAX* as a positional candidate in the aetiology of the t(1;11) phenotype. Although there have been previous investigations into the function of the TRAX protein, its biological roles remain poorly defined. In addition there is evidence for further transcripts in proximity of the human *TRAX* gene, including the *Bactrax* transcript that appears to share its core promoter with *TRAX* (section 5.2.2). The significance of these other transcripts, intergenic splicing with *DISC1* and the associated intergenic exons is unclear. To gain further insight into the function of *TRAX*, its genomic organisation, associated transcripts and regulatory elements, comparative analysis of the *TRAX* gene was undertaken at the genomic and protein level.

8.2 Introduction

Translin, a paralogue of *TRAX*, was identified by Aoki *et al.*, (1995) as a protein that binds recombination hot spots and is associated with chromosomal translocations in human lymphoid malignancies. *TRAX* was first identified in a yeast two-hybrid screen for proteins that interact with translin (Aoki *et al.*, 1997). This interaction was confirmed by an *in vitro* interaction assay (Aoki *et al.*, 1997) and has subsequently been replicated by several other groups (Taira *et al.*, 1998; Finkenstadt *et al.*, 2000; Chennathukuzhi *et al.*, 2001a). As well as a role in the induction and/or repair of double stranded DNA breaks, *TRAX* and Translin both have roles in the

binding and sub-cellular transport of specific mRNA species. Translin remains the better characterised of the two interacting and homologous proteins, for this reason the subjects of mRNA and DNA binding are introduced by the work previously carried out on Translin, the role of *TRAX* is then discussed in this context.

8.2.1 mRNA binding and transport

Translin functions as an RNA binding protein mediating the transport and temporary translational suppression of mRNA species characterised by a conserved pair of sequence motifs, the Y and H elements (Kwon *et al.,* 1991; Kwon *et al.,* 1993; Morales *et al.,* 1998; Kobayashi *et al.,* 1998; Hecht 2000 for review). Sequence similarity searching has identified many transcripts with sequences similar to the Y and H elements recognised by Translin (Han *et al.,* 1995). Specific interactions have been demonstrated between Translin and the mRNAs of protamine 1, protamine 2, *AKAP 82*, mylin basic protein, α-calmodulin kinase II and Tau as well as the apparently non-protein coding BC1 RNA (Wu and Hecht, 2000; Han *et al.,* 1995; Kobayashi *et al.,* 1998).

Translin was also found to bind microtubules, either reassembled *in vitro* from crude brain extract or purified tubulin (Han *et al.,* 1995) and was found to co-localise with microtubules within neurons (Wu and Hecht 2000). Combined, these results demonstrate that Translin has an intrinsic ability to bind microtubules. Han *et al.,* (1995) also showed that Translin bound to microtubules retained the capacity to bind RNA molecules in a sequence specific manner. The binding of Translin to microtubules was found to require the concomitant assembly of microtubules, and was diminished by inhibitors of microtubule assembly (colcemid, cytochalasin D and high salt concentrations) (Han *et al.,* 1995).

The combined properties of mRNA binding, translational suppression and microtubule binding lead to the hypothesis that Translin may be involved in the transport and anchoring of mRNAs to specific sub-cellular locations in a manner akin to the *nanos* and *bicoid* mRNAs of *Drosophila* (van Eeden and St Johnston, 1999 for review). The dendritic translocation of Translin – RNA complexes has been reported in rat hippocampal neurons (Kobayashi *et al.,* 1998) and in human

hippocampal and neocortical pyramidial neurons (Finkenstadt *et al.,* 2000). The Translin mediated transport of mRNAs has also been demonstrated in developing spermatogonia and between spermatogonia via intracellular, cytoplasmic bridges (Morales *et al.,* 1998; Hecht, 2000). These findings indicate that Translin mediates the specific targeting of mRNAs to sub-cellular locations and provides a means of inhibiting translation (by an unknown mechanism) until its delivery. It is also plausible that mRNA translation is inhibited until a specific signal is received to 'unmask' the transcript at its site of delivery, this possibility and its implications are discussed in section 8.7.

8.2.2 Single stranded DNA binding activity

In addition to sequence specific binding of RNA, it has been demonstrated that Translin has the ability to bind single stranded DNA (ssDNA) in a sequence specific manner (Aoki *et al.,* 1995; Kasis *et al.,* 1997 Wu *et al.,* 1998; Chennathukuzhi *et al.,* 2001a). The DNA sequence motif recognised by Translin is a bipartite motif that shares no obvious similarity to the consensus RNA motif that it recognises. Binding is only efficient if the motif occurs near the end (within 12 nucleotides) of the DNA molecule (Kasai *et al.,* 1997).

The DNA binding properties of Translin were first identified in a screen for factors that bound a DNA motif that had been repeatedly found in close proximity to chromosomal translocation breakpoints in lymphoid malignancies (Aoki *et al.,* 1995). The Translin binding motif was present both at breakpoints associated with the heptamer – nonamer sequence motifs recognised the Ig/TCR recombinase and at breakpoints without these characteristic motifs (Bassing *et al.,* 2000 for review). Translin recognition sites have also been found in non-lymphoid, tumour associated translocations (Kanoe *et al.,* 1999; Chalk *et al.,* 1997) arguing that the role of Translin in induction and/or repair of the double stranded DNA breaks is not restricted to lymphoid tissue or an exclusive association with the Ig/TCR somatic recombination system.

8.2.3 TRAX – Translin protein interactions and nucleic acid binding

Translin dimers are the minimum unit for nucleic acid binding; they are able to bind both RNA and ssDNA in a sequence specific manner, as assessed by mobility shift assay (Kasai *et al.,* 1996; Wu *et al.,* 1997; Chennathukuzhi *et al.,* 2001a). Translin has also been observed to form larger aggregate, ring shaped structures (Kasai *et al.,* 1996) that retain the ability to bind DNA (the RNA binding ability of these structures has not been tested). TRAX has been found to interact with Translin by yeast two hybrid assay and co-immunoprecipitation (Aoki *et al.,* 1997; Taira *et al.,* 1998). Using affinity column purification, Taira *et al.,* (1998) have clearly shown by that TRAX is bound to Translin when Translin is bound to DNA, with the two proteins in an approximate 1:1 ratio.

Chennathukuzhi *et al.,* (2001) have shown that TRAX alone is unable to bind DNA or RNA and is unable to homodimerise. Chennathukuzhi *et al.,* (2001) demonstrated that TRAX and Translin were able to form heterodimers as had previously been indicated by yeast two hybrid work (Aoki *et al.,* 1997). TRAX – Translin heterodimers were unable to bind RNA, but the sequence specific binding of ssDNA was enhanced in the heterodimer compared to the Translin homodimer (Chennathukuzhi *et al.,* 2001a).

Investigations into the sub-cellular localisation of TRAX and Translin also reveal clues to the regulation of these proteins and their roles within the cell. Translin was found to be a predominantly cytoplasmic protein in most cells, although it was observed to move to the nucleus upon DNA damage or following VJD recombination in lymphoid cells (Kasai *et al.,* 1997; Chennathukuzhi *et al.,* 2001a). TRAX was predominantly localised to the cytoplasmic fraction by western blot analysis and observed to co-localise with endoplasmic reticulum and Golgi markers by GFP fusion and confocal microscopy (Chennathukuzhi *et al.,* 2001a). It has been proposed that upon DNA damage, TRAX (which contains a predicted nuclear localisation signal) mediates the nuclear transport of Translin (which does not contain a predicted nuclear localisation signal) (Aoki *et al.,* 1997). The sub-cellular location of TRAX following DNA damage has not yet been investigated, although

the induced nuclear localisation of Translin and by inference TRAX upon DNA damage would be consistent with the observed nucleic acid binding properties of the TRAX – Translin heterodimer. Both TRAX and Translin are expressed in all tissues tested, with higher levels of Translin expression in the brain and testes (Devon *et al.,* 2000; Chennathukuzhi *et al.,* 2001a).

8.3 Cloning mouse *TRAX*

The coding sequence of human *TRAX* cDNA (embl:X95073) was searched agsinst the EMBL database using BLASTN (section 2.11.2), identifying mouse TRAX ESTs. Eleven highly homologous ESTs (AU035805, AI663182, AI852243, AI787676, AU035582, AI509075, AU080751, AW227868, AI430863, AA437471 and AA118172) that covered 86% of the expected coding sequence (based on alignment to the human open reading frame) were identified. These ESTs were used to seed the iterative assembly process described in section 2.11.4, resulting in the assembly of a complete open reading frame and 3' UTR of a mouse *TRAX* transcript.

Subsequent work by R. Devon (MRC Human Genetics Unit) confirmed the sequence of mouse *TRAX*, demonstrated its expression by RT-PCR and Nothern analysis and mapped the gene to mouse chromosome 8 in a 5 cM region between markers D8Mit49 and D8Mit280. Localisation to chromosome 8 was confirmed by metaphase FISH analysis (Devon *et al.,* 2000). Thus, mouse *TRAX* maps to a region of known conserved synteny with human chromosome 1q42 (Blake *et al.,* 2001).

The assembly of fragments of the mouse *TRAX* genomic locus from whole genome shotgun sequences (not shown) led to the identification of two previously unknown sequences showing 93% and 94% nucleotide identity with mouse *TRAX* coding sequence. Both of these sequences lack introns, and insertions, deletions and point mutations have disrupted the *TRAX* reading frame, strongly suggesting that both represent processed pseudogenes (data not shown). There is no evidence for the transcription of these pseudogenes in the EST databases. However, a transcript that included all or part of one of the pseudogenes could explain the 2 kb band on mouse

Northern blots hybridised with *TRAX* derived probes (Devon *et al.,* 2000), a band that is not readily explained by the known structure and splicing of the mouse *TRAX* gene. Although not discussed further, a knowledge of such sequences is important in the interpretation of PCR and hybridisation results and may have implications for future strategies in the generation of animal models of the chromosome 1 breakpoint locus.

8.4 Genomic structure of *Fugu TRAX*

8.4.1 *Fugu TRAX* gene structure prediction

Within the *Fugu* sequence contig, all detected similarity (TBLASTN and dotter, section 2.11.2) to human and mouse *TRAX* was contained within the 5 kb *EcoRI* fragment of the *Fugu* genomic sequence contig (section 3.3.4). *Ab initio* gene prediction methods applied specifically to the 5 kb *EcoRI* fragment consistently predicted three exons (figure 8.1), with other exons being predicted with only one or a few of the methods used.

Alignment of human and mouse amino acid sequence against the 5 kb *EcoRI* fragment identified four probable exons of *TRAX* (figure 8.1), consistent with genscan predictions of gene structure. Integrating homology data and gene model prediction (Genewise, section 2.11.2), four well predicted exons with splice sites and phases in the same relative positions as human and mouse *TRAX* were defined (section 8.5.1). However, there was no sequence similarity to human nor mouse exons 1 and 2 within the 5 kb *EcoRI* fragment (figure 8.1) or else were in the *Fugu* sequence contig (data not shown).

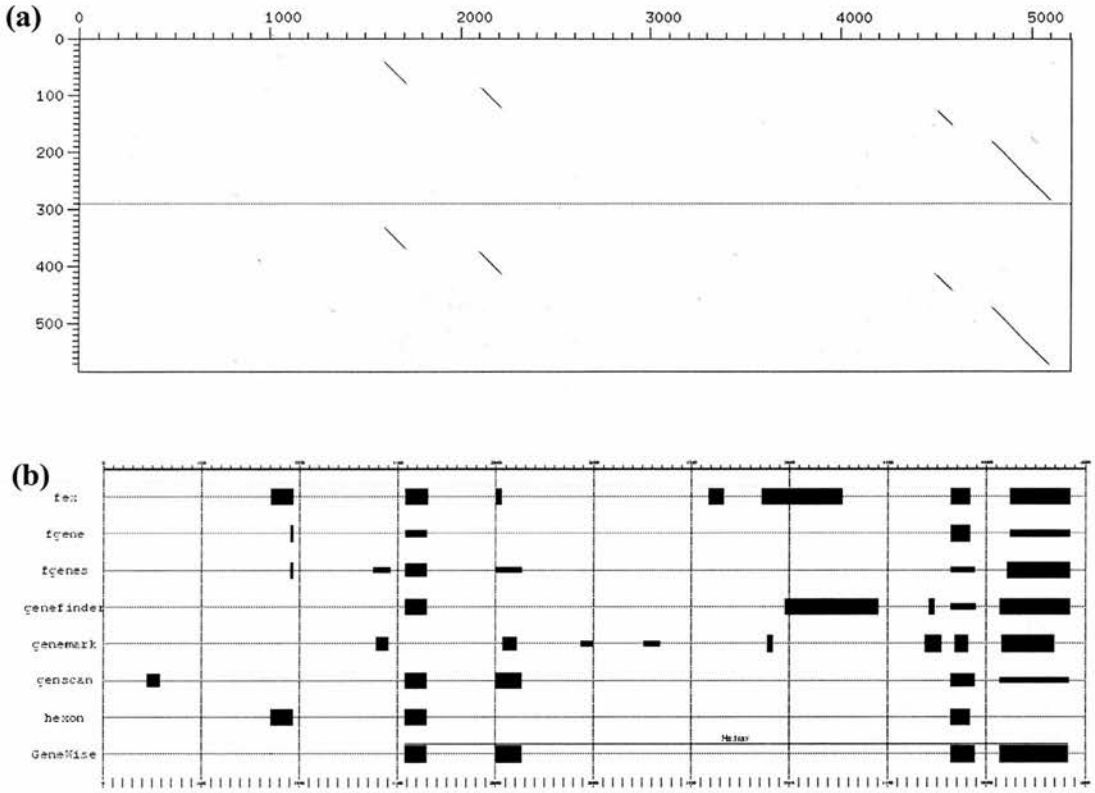


Figure 8.1; *Ab initio* and homology based prediction of *Fugu TRAX*. **(a)** Alignment of human (Hs) and mouse (Mm) *TRAX* amino acid sequence against the 5 kb *Fugu* genomic *TRAX* fragment. Human amino acid sequence is represented by the vertical axis above its intersection by the horizontal line. Mouse amino acid sequence is represented by the vertical axis below the green line. **(b)** Summary of exon prediction for the 5 kb *Fugu* genomic *TRAX* fragment. Only forward strand predictions are shown. *Ab initio* methods fex, fgene, fgenes, genefinder, genemark, genscan and hexon were used along with the homology based GeneWise method (section 2.11.2). GeneWise was provided with the amino acid sequence of human *TRAX* as a basis for homology based gene prediction. All programs were run with default settings. Vertical lines indicate 0.5 kb blocks of sequence.

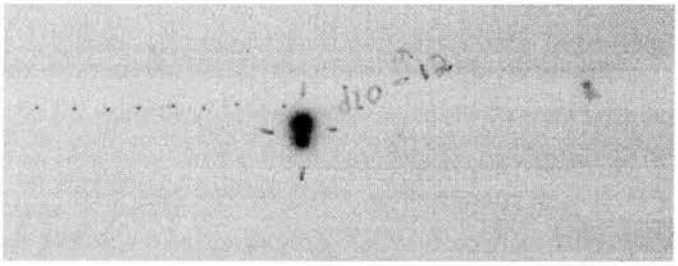
8.4.2 *Fugu TRAX* gene structure confirmation

Based on genscan and Genewise prediction of gene structure (figure 8.1) a 269 bp probe (FT-P1, section 2.10.10) was designed from *Fugu* genomic sequence corresponding to the predicted terminal exon of *TRAX*. Hybridisation screening of 11 *Fugu* cDNA libraries (107,000 clones, section 2.9.7) identified a single gut derived cDNA clone (d10-m12) giving a strong and specific hybridisation signal to the FT-P1 probe (figure 8.2). The insert of clone d10-m12 was completely sequenced to high quality (Phrap score of >30 for every base) demonstrating that it represented a *Fugu TRAX* transcript encompassing the entire open reading frame, full 3' UTR and a stretch of 5' UTR. Alignment of this cDNA clone to the *Fugu* genomic sequence allowed the resolution of *Fugu TRAX* genomic structure (figure 8.2), confirming the Genewise predicted structure of exons 3 to 6 and identifying the previously undetected first and second exons of the gene.

8.4.3 Homology based prediction of *Tetraodon TRAX* gene structure

Iterative sequence clustering and assembly (section 2.11.4) of *Tetraodon nigroviridis* whole genome shotgun sequence resulted in the assembly of 7 kb of contiguous sequences across the *TRAX* genomic locus of *Tetraodon* (section 5.3.2). The experimentally demonstrated genomic structure of *Fugu TRAX* and the relatively close evolutionary relationship of *Tetraodon* and *Fugu,* allowed the *Tetraodon* protein coding sequence to be annotated with reasonable confidence through Genewise (section 2.11.2) alignment with the *Fugu TRAX* amino acid sequence (figure 8.4). With the exception of intron 1 (section 8.5.1), the gene structure of *Tetraodon TRAX* was predicted to be the same as that of the *Fugu* orthologue.

(a)



(b)

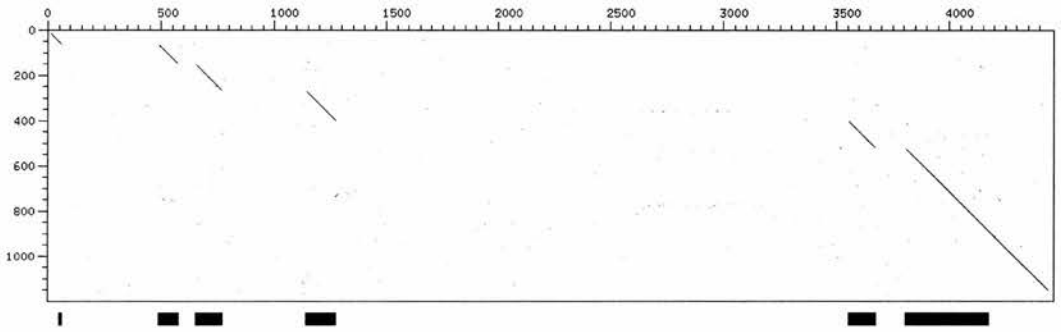


Figure 8.2; *Fugu TRAX* transcript and genomic structure. **(a)** Probe FT-P1 screened against a gridded, double spotted *Fugu* gut derived cDNA library, identifying a single positive clone d10-m12. **(b)** The full insert sequence of clone d10-m12 aligned against *Fugu* genomic sequence (coordinates 8101 to 12550 of the sequence reported in section 3.5.6). Black bars under the dotter alignment show the position of protein coding sequence in the alignment.

8.5 Evolutionary conservation of the *TRAX* gene

8.5.1 Gene structure conservation

For each of the orthologous human, mouse, *Fugu* and *Tetraodon* genes, *TRAX* comprises six exons with each exon contributing to the open reading frame. The relative position and splicing phase of introns 2 to 5 was found to be conserved between human, mouse, *Tetraodon* and *Fugu TRAX* genes (figure 8.3). The splicing of intron 1 was found to be phase one in human and mouse and homology based prediction of gene structure also predicted *Tetraodon* intron 1 as phase one (figure 8.4). However, if a phase one intron was assumed, *Fugu* genomic sequence around the intron 1 splice donor and acceptor sites did not conform to the normal vertebrate splice site consensus sequence (figure 8.4). If the splice donor and acceptor sites were shifted by one residue and a phase two intron was assumed, the splice donor site then conformed to the consensus sequence, although the splice acceptor remained a poor match (figure 8.4).

From the data available, it was not possible to determine if the *Fugu* intron has shifted in splice phase or if it is utilising highly unusual splice sites. It has previously been noted that only 0.02 % vertebrate splice sites do not conform to the previously described families of splice sites (Burset *et al.,* 2000). If splicing was assumed to be phase one, then the AG-CG combination of splice site dinucleotides proposed for *Fugu TRAX* intron 1 would provide the first description of this non-canononical splice site dinucleotide pair in vertebrates (Burset *et al.,* 2000). On the basis of this evidence it is most likely that intron 'sliding' has occurred in the *Fugu* lineage since the last common ancestor with *Tetraodon*, resulting in the phase shift of *Fugu* intron 1. The caveat to this argument is that vertebrate spice site consensus sequences have been determined almost entirely on mouse and human data sets, *Fugu* and fish in general may have different constraints on splice site usage, although this has not been reported for the fish genes analysed to date.

For each of the orthologous genes, the first exon consists of mostly 5' UTR and 17 nucleotides (18 in the case of *Fugu*) of the TRAX open reading frame. The

predicted translation initiation sites for mouse, human, *Fugu* and *Tetraodon TRAX* conform well to the Kozak consensus sequence (Kozak, 1996). Two further lines of evidence also strongly support the predicted site of translation initiation: (a) The absence of upstream ATG codons in any of the three forward frames of mouse, human or *Fugu* transcripts, consistent with the "linear scanning" model of translation initiation (Cigan *et al.,* 1988). (b) The presence of in-frame stop codons upstream of the predicted translation initiation site in mouse and *Fugu.*

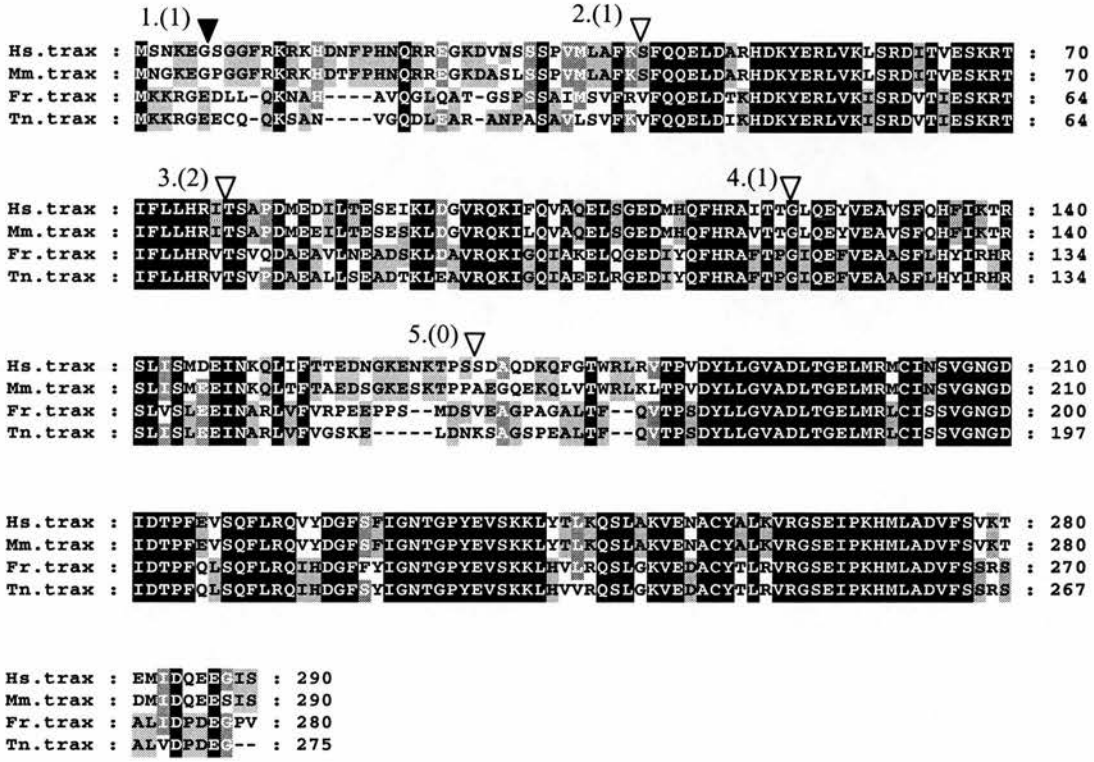


Figure 8.3; The conserved genomic structure of vertebrate *TRAX* genes. Multiple sequence alignment of human (Hs), mouse (Mm), *Fugu* (Fr) and *Tetraodon* (Tn) *TRAX* amino acid sequences. Black, dark grey and light grey backgrounds for 100%, 66% and 50% residue identity respectively. Triangles indicate the relative position of introns, an open triangle indicating conservation in each of the aligned genes. The Black triangle indicates intron 1 which may be phase 2 in *Fugu*. Introns are numbered 1 to 5, numbers in parenthesises indicate the phase of the intron relative to the reading frame.

Figure 8.4; Intron 1 splice sites – evidence for intron sliding. **(a)** Consensus splice sites found in vertebrate introns (Moore, 2000). Exon residues indicated in capital letters and the highly conserved intron boundary residues shown in bold. **(b – e)** Alignments of splice donor sites (green) of exon 1 and splice acceptor sites of exon 2 from **(b)** *Tetraodon*, **(c)** *Fugu*, **(d)** human and **(e)** mouse. The start codon is highlighted in bold and italics for each alignment. Splice donor (green) and acceptor (red) sites were aligned with the experimentally determined cDNA sequences (cDNA was not available for *Tetraodon*). Intron residues indicated by lower case, exon residues indicated by upper case. Residues that align between the splice donor, splice acceptor and cDNA sequence indicated in bold, highlighting these regions of alignment ambiguity. Splice site consensus sequences were used to predict the splice site within the sort region of ambiguity. The relative position of the predicted splice site was found to be conserved between *Tetraodon*, human and mouse but has moved by one residue in *Fugu*, altering the phase of the splice site.

8.5.2 Intron size conservation

Multi-exon *Fugu* and *Tetraodon* genes are typically eight to ten times smaller than their human orthologues, reflecting the compact nature of these vertebrate genomes (Elgar *et al.,* 1999 and section 1.7.1). Exons typically remain a similar size between species, whereas introns are dramatically reduced in the compact genomes (section 1.7.1). Although it has not yet been commented on in the literature, there appears to be a good correlation in the rank ordering of intron size for orthologous genes between vertebrate species (table 8.1 and personal observation). It is not known if this correlation reflects functional constraint on specific introns or the mechanisms of genome compaction and expansion.

Evaluation of *TRAX* intron sizes between *Fugu*, *Tetraodon* and human finds a perfect rank order correlation between *Fugu* and *Tetraodon* with intron 2 being the smallest followed by 5, 3, 1 and intron 4 being the largest intron. For human *TRAX*, this pattern is disrupted slightly, most noticeably by intron 1, which is only 3 bp larger than *Fugu* intron 1 and 52 bp smaller than *Tetraodon* intron 1. The conservation of intron 1 length may reflect specific functional constraint on the intron, either in terms of its length or the sequence contained within it.

It is interesting to note that the ratio of *TRAX* intron sizes between humans and *Tetraodoniformes* varies from 0.9 to 103, yet the ratio of total gene length is in close concordance with an average 8 to 10 fold genome and gene length compaction.

(a) Intron length

Intron	1	2	3	4	5	Total length
Human	427	7869	5158	18520	3273	35115
Fugu	424	76	371	2278	132	4168
Tetraodon	479	82	333	2062	96	3889

(b) Size ratios

Human : *Fugu*	1.0	103	13.9	8.1	24.8	8.4
Human : *Tetraodon*	0.9	96.0	15.5	9.0	34.1	9.0

Table 8.1; Intron size conservation. **(a)** The length of *TRAX* introns as defined by alignment of cDNA sequence to genomic sequence (or homology based gene prediction in the case of *Tetraodon*), using splice site consensus sequences to resolve alignment discrepancies. Total length reflects the total gene length as defined by stop codon coordinate minus start codon coordinate. Units are in base pairs. **(b)** Size ratios of human versus *Fugu* and *Tetraodon* intron length. Total gene length is defined as the distance between the start codon and stop codon in genomic DNA. The ratio is calculated as: human length / fish length.

8.5.2 Non-coding homology between the *TRAX* transcripts

There is no detectable sequence similarity between the 5' UTR of *Fugu* and mammalian (human and mouse) *TRAX* transcripts. Within the 3' UTR, mouse and human share at least four blocks of conservation above the background level (figure 8.5). Similarly, there is a striking block of 36 residues sharing 86% identity in un-gapped alignment between the 3' UTR of *Fugu TRAX* and sequence down stream of the *Tetraodon TRAX* gene, which was also predicted to be located within the 3' UTR (figure 8.5). Genomic sequence alignment indicated that the block of *Tetraodon* : *Fugu* conserved 3' UTR sequence (figure 8.5) corresponded to one of the four blocks conserved between humans and mice (figure 8.5). Manual alignment of the 3' UTR conserved nucleotides, identified a short (41 to 43 nucleotides) and highly conserved sequence motif present towards the 3' end of the 3' UTR in each species. An interesting possibility was that these conserved regions may contain the Y and H element sequences recognised by the *TRAX* interacting protein Translin (Han *et al.,* 1995; Muramatsu *et al.,* 1998) providing a potential means for auto-regulation. However, no significant similarity (as defined by Han *et al.,* 1995) was

found between the Y or H motifs in the highly conserved or mammalian specific 3' UTR sequences. Using the criteria of Han *et al.,* (1995), the only similarity in the *TRAX* transcripts of *Fugu*, *Tetraodon*, mouse or human to either of these motifs was a match to the Y- motif in the 3' UTR of mouse *TRAX* (bases 2241 – 2254 of the reference sequence, Apendix I). This similarity was in a region that was not conserved between mouse and any of the other species investigated (data not shown).

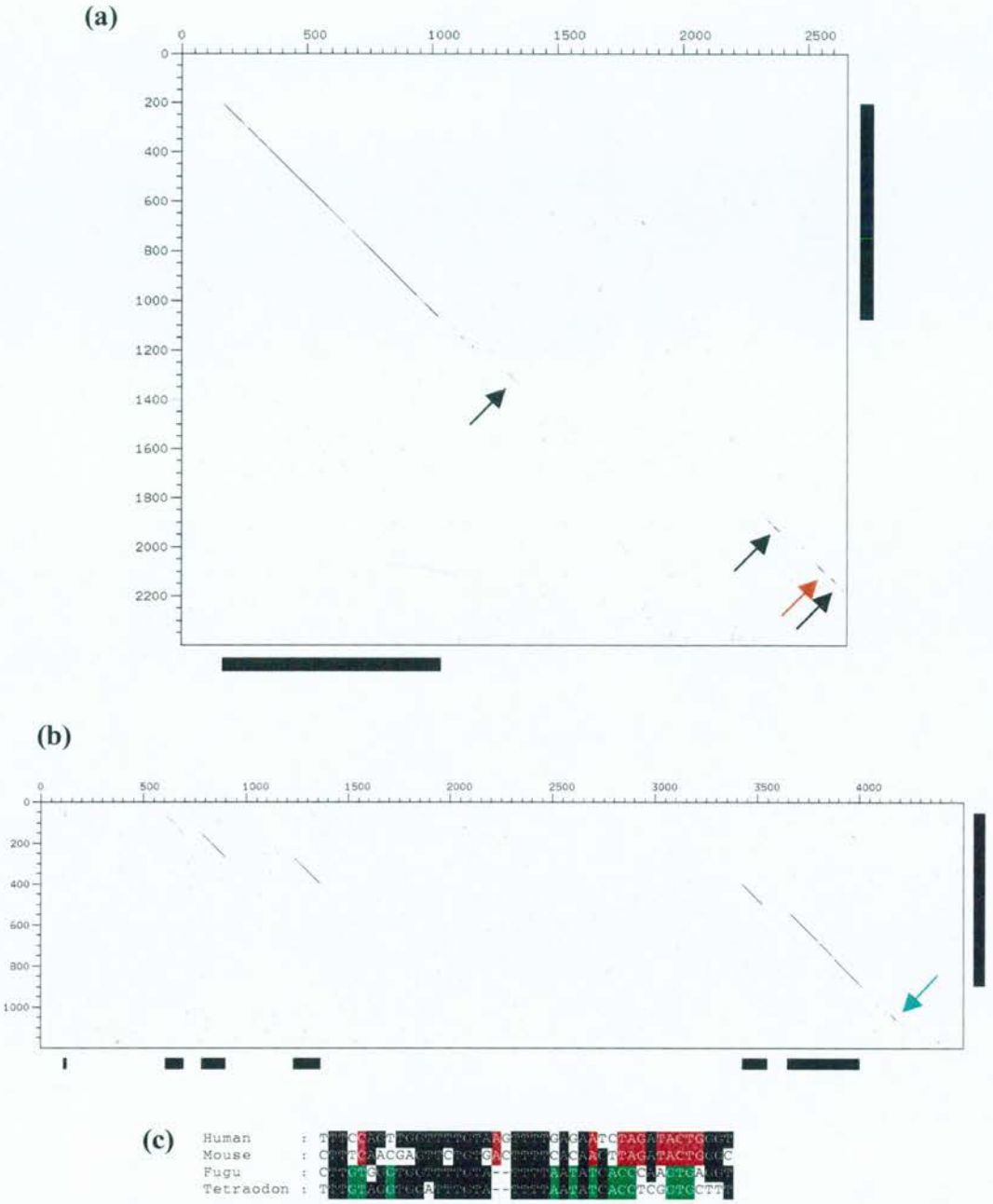


Figure 8.5; Conserved non-coding sequences in the 3' UTR of *TRAX* transcripts. **(a)** Dotter alignment of human (horizontal) and mouse (vertical) *TRAX* transcripts. Coding sequence is indicated by black bars along the respective axis. Blocks of conserved non-coding sequence are indicated with arrows. The brown arrow indicates sequence shown in panel 'c'. **(b)** Dotter alignment of *Tetraodon* genomic sequence (horizontal) containing the *TRAX* gene and the *Fugu TRAX* transcript (vertical). Black bars along the respective axis indicate coding sequence. The single block of conservation between *Fugu TRAX* 3' UTR and the

equivalent genomic sequence from *Tetraodon* is indicated with an arrow, the sequence of which is shown in panel 'c'. **(c)** Multiple sequence alignment of conserved 3' UTR sequences identified in human, mouse, *Fugu* and *Tetraodon TRAX* genes. Nucleotides identical only between mammalian species are highlighted with a brown background, those identical only between *Tetrodontiformes* highlighted with a green background and those conserved (>75%) between groups indicated with a black background.

8.5.4 Comparative genomic alignment of the *TRAX* gene region

The assembly of contiguous *Tetraodon* sequence over the *TRAX* genomic region from 110 bp upstream of the predicted start codon to downstream of the stop codon provided a valuable additional resource for comparative genomic analysis of the region. For this reason, analysis of the structural gene region was separated from the comparative analysis of the promoter and upstream sequences for which there was interspersed *Tetraodon* sequence coverage.

Genomic sequence alignment (Avid, section 2.11.2) between *Fugu* and *Tetraodon* sequence shows greater than 85% sequence identity in *TRAX* exons 3, 4, 5 and 6; 67% identity for exon 2; and poor conservation of exon 1 (figure 8.6). Within apparently non-coding sequences, there was a dramatic variation in the distribution of sequence conservation, ranging from no detectable similarity to a 652 bp block with 80% identity in intron 4 (figure 8.6). From this genomic alignment alone, it would be easy to draw the conclusion that there had been sufficient divergence of sequences without functional constraint, that all observed sequence conservation between orthologous *Fugu* and *Tetraodon* regions reflected functional constraint. Closer examination of each non-similar region reveals specific reasons for the lack of conservation. Intron 3 is very similar in length in both *Fugu* and *Tetraodon* (371 and 333 bp respectively), but shares little sequence similarity (figure 8.6). Searching the *Fugu TRAX* intron 3 against *Fugu* genomic survey sequence (Elgar *et al.,* 1999) and whole genome shotgun sequence has identified a novel interspersed repetitive element that appears specific to the *Fugu* genome with at least 150 copies in the genome (section 5.3.1). At least 166 nucleotides of *Fugu TRAX* intron 3 are accounted for by this repetitive element that is not present in the *Tetraodon TRAX* region. The approximately 600 bp region of *Fugu* intron 4 that has no detectable

homology with the *Tretraodon* sequence is accounted for by 12 tandem copies of a novel 48 nucleotide sequence (figure 8.6).

Even though the background level of conservation may have been underestimated in the genomic sequence alignments between *Fugu* and *Tetraodon* sequences, the conservation of 652 bp at 80% identity and 204 bp at 76% identity within intron 4 was substantially above the background level of conservation (see figure 5.5 comparison) that presumably occurs by chance alone.

As expected, the overall level of sequence similarity between fish and human sequences was substantially less than that ovserved between more colsely related species (figure 8.7). Exons 3, 4, 5 and 6 showed >65% identity with reduced, but detectable, sequence similarity for exons 1 and 2 (figure 8.7).

There was a cluster of sequence conservation within the first intron of the *TRAX* gene with the whole intron showing 40 to 60% identity in alignment between fish and humans and 60 to 75% identity between *Fugu* and *Tetraodon* (figure 8.7 and 8.8). This conservation along the entire length of the intron and the unusually consistent size of the intron between species (table 8.1) suggests that the first intron may contain, or be largely composed of, transcriptional or other regulatory elements. The specific absence of repetitive elements has previously been assocated with sites of regulatory importance (The International Human Genome Sequencing Consortium, 2001). The observation that human intron 1 is completely devoid of repetitive elements also supports the hypothesis that vertebrate *TRAX* intron 1 harbours or is largely composed of regulatory sequences.

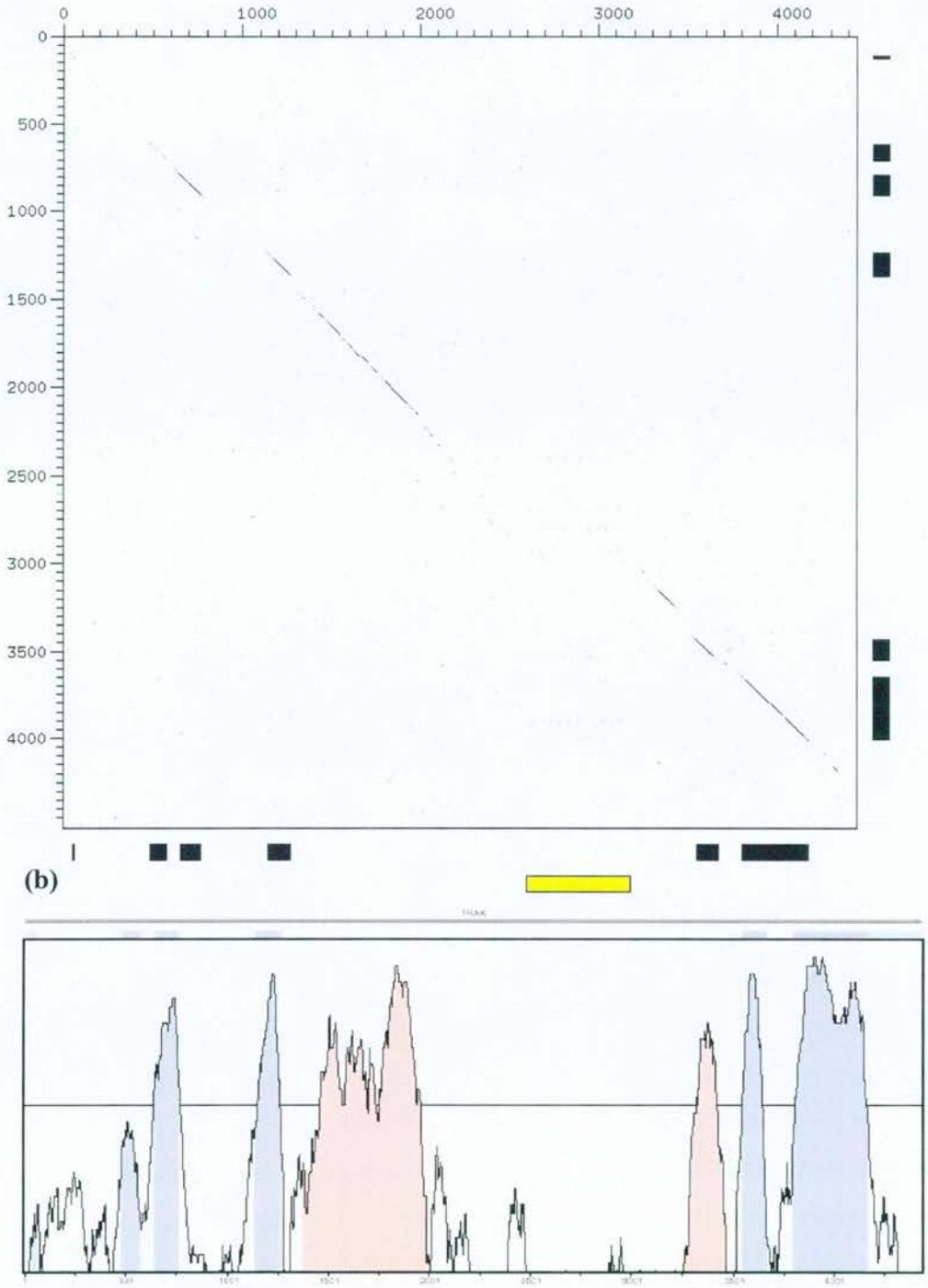


Figure 8.6; *Fugu – Tetraodon* comparative genomic alignment. **(a)** Dotter alignment (window size 22) of the *Fugu TRAX* genomic region (horizontal) and the orthologous region from *Tetraodon* (vertical). Black boxes along each axis indicate coding sequence of the *TRAX* gene. The yellow box along the horizontal axis highlights a region of tandem 48 nucleotide repeats specific to *Fugu*. **(b)** Vista plot of the *Fugu TRAX* genomic region and orthologous region from *Tetraodon* (the same sequences as shown in panel 'a'). The horizontal axis represents the *Fugu* sequence and vertical axis the percent identity in global

alignment (Avid, section 2.11.2) with the *Tetraodon* sequence, averaged in a sliding window over 40 residues.

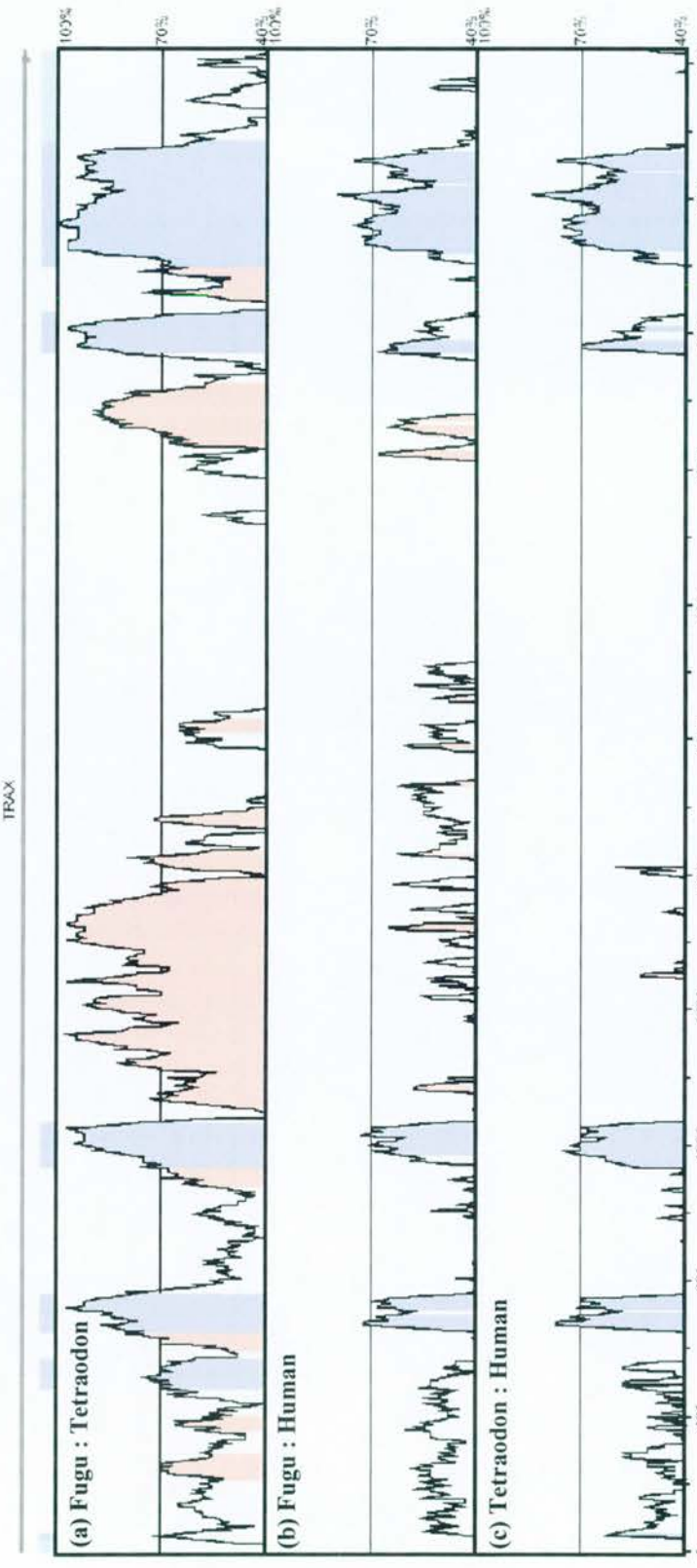


Figure 8.7; Global sequence alignment of the *Fugu*, *Tetraodon* and human TRAX genomic regions. Vista representation of sequence alignment (sections 2.11.2 and 5.5) using experimentally demonstrated exon splice sites (human and *Fugu*) and predicted splice sites (*Tetraodon*) as cross-species anchors for Avid alignment (section 2.11.2). The horizontal axis shows the scale for *Fugu* genomic sequence. Coding sequence is indicated by the dark blue boxes above the graph, 5' and 3' UTR by pale blue boxes. Graphs plot average identity in the alignment averaged over a sliding window of 40 nucleotides, regions above 66% identity are coloured pink or reflect the colour of the feature they represent (coding or UTR). **(a)** Alignment between *Fugu* and *Tetraodon*. **(b)** Alignment between *Fugu* and human. **(c)** Alignment between *Tetraodon* and human, coordinates of the alignment were transformed (Avid, section 2.11.2) to equivalent *Fugu* coordinates to enable comparison of alignments.

8.5.5 Comparative promoter analysis

The core promoter elements for polymerase II transcribed genes are generally located directly or a short distance upstream of the transcription start site (Watson *et al.,* 1987). Identification of the *TRAX* promoter as a bi-directional promoter (section 5.2.2) allows the core of this promoter to be characterised using the flanking transcriptional start sites as approximate boundaries for the analysis, although other regulatory elements are likely to be located outside this region.

No sequence similarity was detected between the human *TRAX* core promoter and its flanking 200 bp and the sequence between the start codon of *Fugu TRAX* and the start codon of the upstream gene, *EGLN1* (data not shown). *Ab initio* promoter prediction (section 2.11.2) did not predict promoters directly upstream of *TRAX* in either *Fugu* or *Tetraodon*. The TSSW promoter prediction program (section 2.11.2) predicted a marginal promoter within the core promoter region of the human gene as defined above. The motifs identified by TSSW in the human promoter were not present in either *Fugu* or *Tetraodon* sequences. None of the three species had a TATA box motif within 100 bp of the expected transcriptional start site.

8.5.6 *TRAX – DISC1* intergenic splicing

Intergenic splicing has been reported between *TRAX* and *DISC1* in the human, although it could only be detected by RT-PCR based methods, indicating that it may be a very rare event (Millar *et al.,* 2000a; see section 5.2.2). Splicing was observed to occur from either exon 4 or exon 5 of *TRAX* into exon 2 of *DISC1* with the variable inclusion of four intergenic exons (A to D) (Millar *et al.,* 2000). Similarly, intergenic splicing has been observed between *TRAX* and *DISC1* in the mouse, although intergenic exons were not used in the mouse transcripts. Again the transcripts were only detectable by RT-PCR based methods. (R. Devon personal communication). The reading frame was not continued between *TRAX* and *DISC1* protein coding reading frames in any of the mouse or human intergenic transcripts. The intergenic exons either introduce a stop codon prior to sequence encoded by

DISC1 exon 2, or *DISC1* exon 2 was spliced to in the wrong phase, causing the reading frame to be shifted, introducing a premature stop codon.

Although the phenomenon of intergenic splicing is conserved between human and mouse, the splicing pattern is not conserved. It is conceivable that the intergenic splicing is functionally relevant, not for protein coding, but rather in the transcriptional regulation of *TRAX* and/or *DISC1*. If this was the case, the lack of a continuous reading frame in the intergenic transcripts would be inconsequential.

Avid nucleotide alignment (section 2.11.2), TBLASTN and TBLASTX sequence similarity searching and dotter alignment of six frame translations of each intergenic exon from human failed to find any convincing sequence similarity with the *Fugu TRAX – DISC1* intergenic region (figures 8.8, 8.9 and data not shown). To test directly for intergenic splicing events, RT-PCR was carried out between *Fugu TRAX* exons 4 and 5 and *DISC1* exon 2. The reactions used *Fugu* heart and ovary cDNA (the only *Fugu* cDNA available). Both samples had previously been demonstrated to express *DISC1* and *TRAX* (sections 6.3 and data not shown). Intergenic splicing could not be detected between *TRAX* and *DISC1* by this method. A control PCR on total *Fugu* genomic DNA and cosmid DNA produced the expected 4 and 4.5 kb bands demonstrating that the PCR is robust (data not shown).

8.5.7 Conserved intergenic sequences

Little sequence similarity was detected between the 100 kb of human genomic sequence and less than 2 kb of *Fugu* sequence that represents the intergenic region between the 5' end of *EGLN1* and the 5' end of *TRAX* (figure 8.8). The notable exception to this is a block of 91 nucleotides in the human sequence that aligned with 68 nucleotides of *Fugu* sequence, showing 61% identity over the whole alignment, but containing blocks of homology within the alignment showing 100% identity for up to 12 nucleotides. Although there was an uninterrupted reading frame for both the *Fugu* and human aligned sequences, the presence of multiple gaps in the alignment that were not multiples of three argues that this conservation did not represent coding sequence. Screening the alignment against the Transfac database (section 2.11.1) using the R-Vista software (section 2.11.2) identified only one

conserved transcription factor binding site within the alignment (Transfac:R00079). The high probability of such a short (5 bp) motif occurring in an alignment of this length and the absence of other motifs suggested that this match was not significant.

There was a noticeable absence of sequence similarity between Human and *Fugu* sequence overlapping the *Backtrax* exons (figure 8.8, see also section 5.2.2 for the *Backtrax* transcript). In humans the *Backtrax* transcripts clearly represent spliced and polyadenylated transcripts, but their protein coding status and significance was unknown (section 5.2.2). A more detailed search for homology was therefore carried out. Six frame translations of each of the Backtrax transcripts were aligned using dotter (section 2.11.2) to the 2 kb *EGLN1 – TRAX* intergenic space using window sizes of 16 to 30. No sequence similarity suggestive of the conservation of a Backtrax like transcript upstream of the *TRAX* gene was detected.

The *Foretrax* transcript (section 5.2.2) located in the intergenic space between *TRAX* and *DISC1* was investigated in the manner described for *Backtrax*. No sequence similarity or evidence from exon predictions indicated that there was a protein coding gene located between *TRAX* and *DISC1*. It is therefore considered unlikely that *Backtrax* or *Foretrax* represent protein coding genes.

As previously shown, *TRAX* intron 1 was found to be conserved between human and *Fugu*, as were blocks of sequence within intron 4 of the gene (section 8.5.4). Downstream of the *TRAX* gene, five regions of the human genomic sequence aligned with the equivalent *Fugu* sequence to above 60% identity over more than 40 nucleotides (figure 8.8).

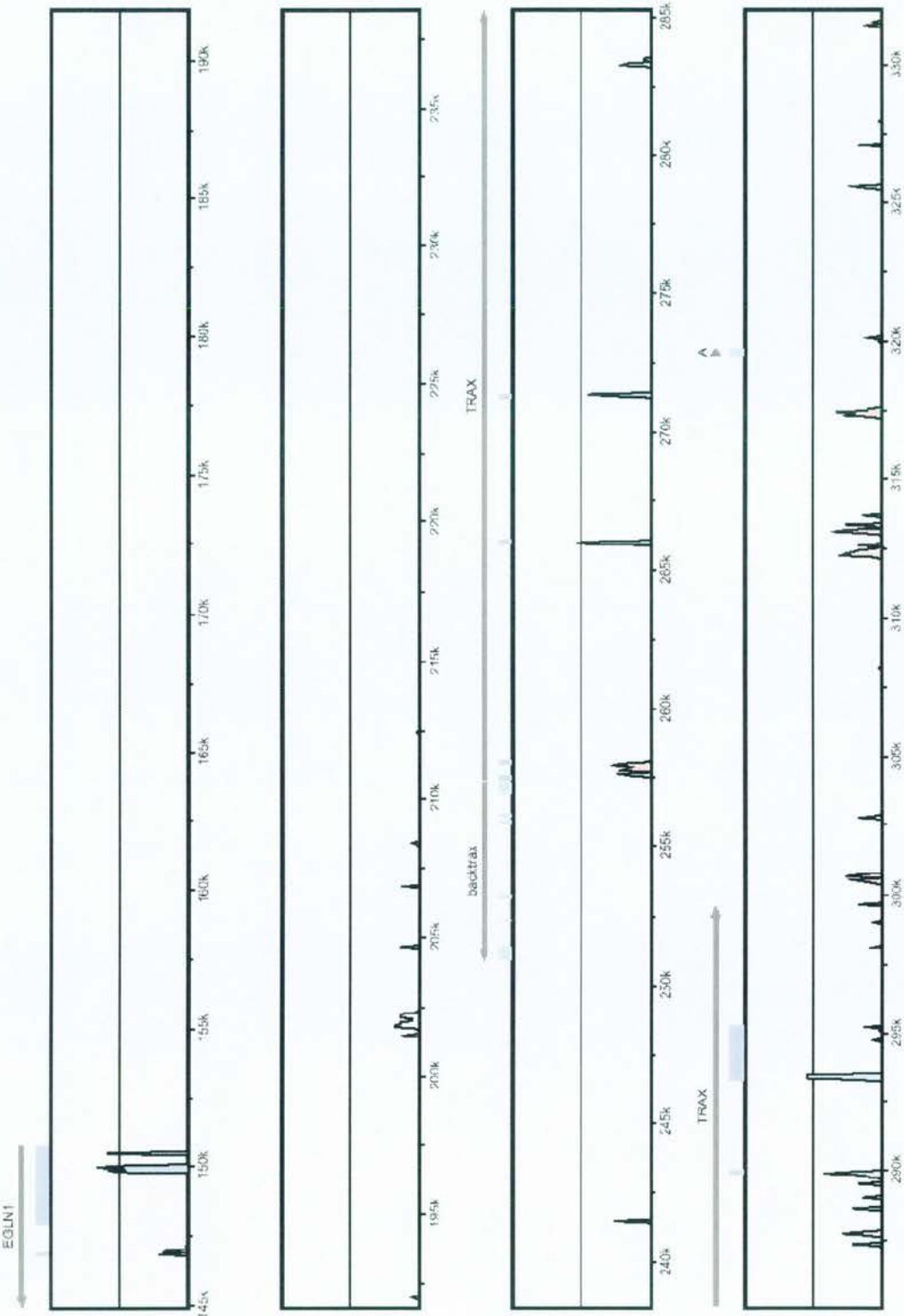


Figure 8.8; Human – *Fugu* genomic sequence alignment. For full legend see next page.

Figure 8.8; Human – *Fugu* genomic sequence alignment. Avid alignment of contiguous human and Fugu genomic sequence using *EGLN1* and *TRAX* exons as annotation anchors (section 5.5). The *TRAX*, Backtrax and *EGLN1* transcripts are annotated and the direction of transcription indicated (grey arrows). The lone exon annotated 'A' indicates a *TRAX* – *DISC1* intergenic exon. Known protein coding exons are indicated in dark blue, transcripts for which the protein coding status is unknown are indicated in light blue. The coordinates along the horizontal axis refer to the assembled human genomic sequence over the *EGLN1* – *TRAX* region (section 4.3). The horizontal line through the plot indicates 70% sequence identity in a sliding window of 40 nucleotides along the alignment. Sequences showing 66% or greater identity in alignment are coloured pink or the colour of an annotated feature that they represent.

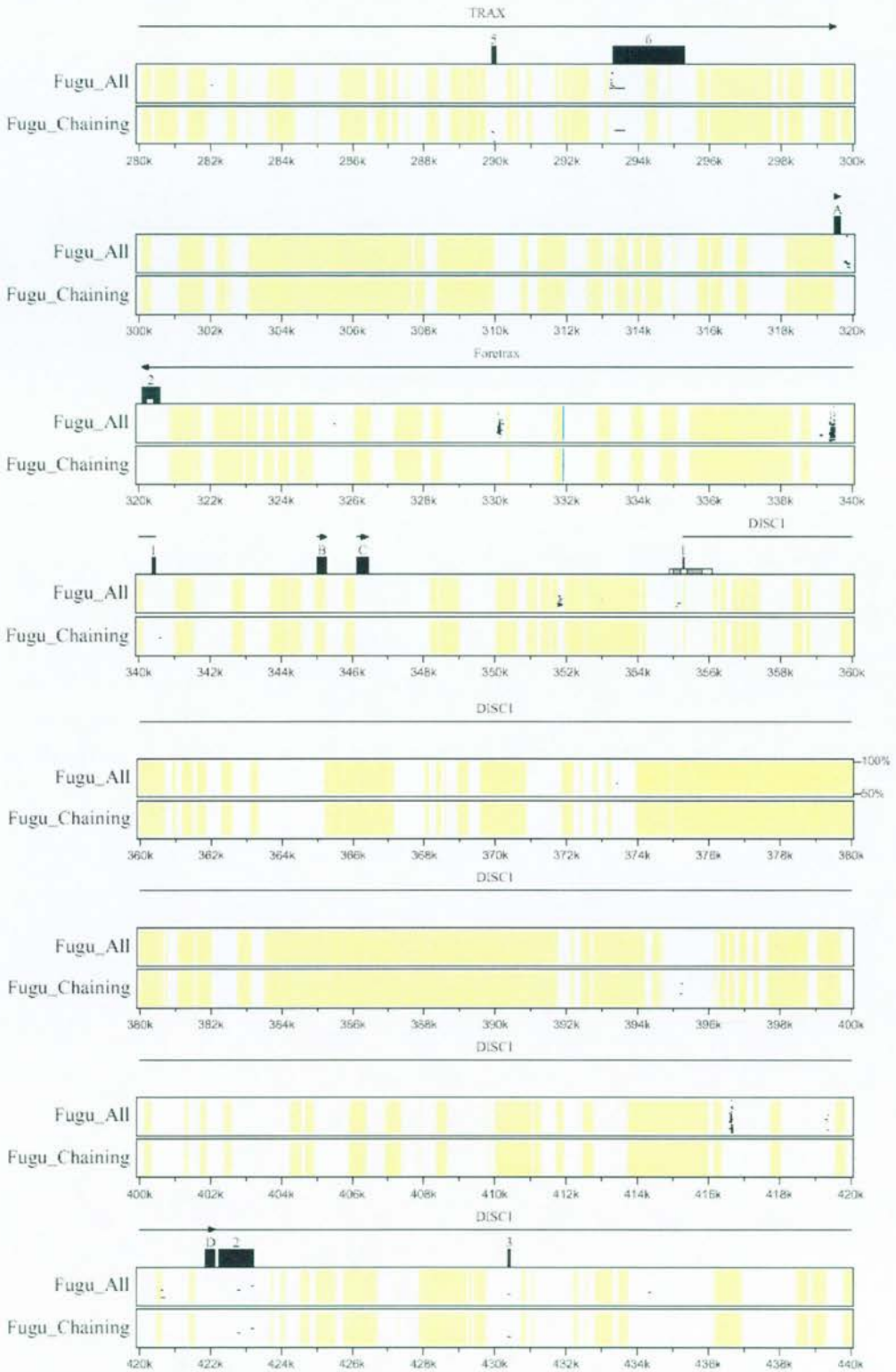


Figure 8.9; Percentage identity plot of the *TRAX – DISC1* intergenic region. See next page of full legend.

Figure 8.9; Percentage identity plot of the *TRAX – DISC1* intergenic region. Percentage identity plot generated with Pipmaker (section 2.11.2). 50 to 100% identity shown for each track of the graph. The base sequence for the alignment is assembled human genomic sequence and the *Fugu* genomic sequence contig is the aligned sequence. The dark blue vertical line at 332 kb is the small gap between assembly fragments (section 4.3). The distal end of the *TRAX* gene and proximal end of the *DISC1* gene are indicated. The transcriptional orientation of annotated features is indicated by a horizontal arrow. *TRAX – DISC1* intergenic exons are indicated by the letters 'A' to 'D'. The track labelled Fugu_All indicates all BLASTZ high scoring segment pairs. The track labelled Fugu_Chaining shows BLASTZ high scoring segment pairs filtered using a chaining model (Pipmaker, section 2.11.2). Human interspersed repetitive elements are indicated by a yellow background to the plot.

8.6 The Translin – TRAX protein family

TRAX and its interacting protein Translin are homologues (33% identity in humans) that have adopted discrete, but related, functional roles within the cell. It is apparent that they are involved in the post-transcriptional regulation and sub-cellular localisation of mRNA (section 8.2.1) and in the induction and/or repair of double stranded DNA breaks (section 8.2.2). However, little is currently known how these functions relate to each other (if at all), the structural basis of the interactions or the mechanistic nature of these roles in the cell. The identification of novel homologous sequences from a wide range of organisms provided a means to investigate some of these issues.

8.6.1 Sequence similarity searches

Searching (BLASTP) the SPTR protein database with the human TRAX amino acid sequence identified TRAX and Translin homologues in metazoans, yeast (*S. pombe*) and plants (*A. thaliana)*. Using TBLASTN to search TRAX and Translin amino acid sequences against EST and genomic sequence databases identified homologues of these genes in many animals and plants (Figure 8.13 and data not shown). Interestingly, orthologues of both *TRAX* and *Translin* are evident in most eukaryotes for which there is substantial genomic or EST sequence available (human, mouse, rat, *X. leveasis*, *T. nigroviridis*, *D. melanogaster*, *S. pombe*, *A. thaliana*, *O. sativa*, *T. aestivum* and *L. esculentum*); (figure 8.10, figure 8.11 and data not shown). This

finding suggested *TRAX* and *Translin* genes diverged from a common ancestor before, or very early, in the eukaryotic lineage.

Neither *TRAX* nor *Translin* homologues were detected in the finished genomes of *C. elegans* or the budding yeast *S. cerevisiae*. Such observations should be treated with caution, as it remains a possibility that there is sequence missing from these genomes. However, the absence of these genes in *S. cerevisiae* may indicate that Translin and *TRAX* originated in the eukaryotic lineage after the divergence of budding yeast but prior to the divergence of fision yeast, animals and plants from each other. The apparent absence of both *TRAX* and Translin from *C. elegans* indicates lineage specific gene loss, and possibly the loss of an entire pathway given that TRAX and Translin are interacting proteins.

8.6.2 *Archaea* bacterial homologues of *TRAX* and *Translin*

Using any of the identified TRAX protein sequences (figure 8.10) as queries in BLASTP similarity searching of the SPTR database (section 2.11.1) readily identified the other *TRAX* proteins as well as the paralogous Translin proteins. Six *Archaea* bacterial predicted proteins were also consistently identified when TRAX proteins were used as a query, but not when Translin proteins were the query. The same region of TRAX protein sequence aligned with both Translin and *Archaea* proteins, ruling out the possibility that separate domains of the TRAX protein were responsible for the discrepancy in detection of similarity between Translin and the *Archaea* proteins.

Psi-BLAST searching (section 2.11.2) of the SPTR database (section 2.11.1), using any of the TRAX, Translin or *Archaea* proteins as the initial query results in the clustering of all three groups of proteins by the second Psi-BLAST iteration. Each of the TRAX, Translin and *Archaea* proteins are clearly homologous within their groups (figures 8.10, 8.11 and 8.12 respectively). The BLAST and Psi-BLAST results suggested that the three groups of protein are homologous and represent a common family of proteins or at least share a common domain. Multiple sequence alignment of TRAX, Translin and the six *Archaea* proteins demonstrates a conserved region covering almost the entire length of each of the homologous

proteins (figure 8.13). This conserved region is subsequently referred to as the Translin domain.

8.6.3 Phylogenetic analysis

Phylogenetic analysis of the aligned Translin domain (figure 8.14) indicates that TRAX and the *Archaea* proteins are more closely related than either is to Translin, as was indicated by the discrepancies in detecting homology to the *Archaea* proteins (section 8.6.1). This suggests one of three probable evolutionary histories for this family of homologous proteins: (a) duplication prior to the divergence of eukaryotes from *Archaea* bacteria and the loss of Translin from *Archaea* bacteria, (b) duplication within the eukaryotic lineage with the subsequent diversification of Translin, (c) horizontal transfer of *TRAX* from eukaryotes to *Archaea* bacteria.

8.6.4 Conservation of sequence motifs

A C-terminal leucine zipper motif and a centrally located basic amino acid motif (figure 8.10) have previously been identified in human Translin (Aoki *et al.,* 1999). The normal homo-dimerisation of Translin was found to be abolished by mutational disruption of the C-terminal leucine zipper. The leucine zipper motif was well conserved between Translin orthologues (figure 8.10) and the homologous region of TRAX and *Archaea* paralogues also shows sequence conservation (figure 8.12). However, the relative insertion of amino acids in the TRAX and *Archaea* proteins would disrupt the heptad periodicity of hydrophobic residues necessary for the formation of a coiled coil. This observation indicates that if Translin does dimerise through a C-terminal coiled coil interaction as proposed (Aoki *et al.,* 1999), TRAX would not be able to heterodimerise with Translin in the same manner.

It has been proposed that the bipartite nuclear localisation signal predicted in the N-terminal region of the human and mouse TRAX protein mediates the conditional nuclear localisation of TRAX and Translin (Kasai *et al.,* 1997). The poor conservation and absence of this sequence from several orthologues (figure 8.10) suggest that it does not have an evolutionarily conserved function.

Two basic amino acid motifs have been identified by mutational studies in Translin as essential for nucleic acid binding (Aoki *et al.,* 1999; Chennathukuzhi *et al.,* 2001a). The N-terminal most motif was essential for DNA but not RNA binding (Aoki *et al.,* 1999) and the C-terminal most motif was essential for both DNA and RNA binding (Chennathukuzhi *et al.,* 2001a) (figure 8.10). Short basic motifs are often involved in binding nucleic acids (Ledent and Vervoort, 2001). However, neither of the motifs were well conserved between orthologues (figure 8.10) nor were the homologous positions always occupied by basic residues. Neither of the basic motifs were conserved in the *Archaea* or TRAX proteins. This lack of conservation suggests that the basic residues are not directly interacting with nucleic acids as suggested by Aoki *et al.,* (1999) and Chennatukuzhi *et al.,* (2001a). It is possible that the mutational disruption of the results affects the protein structure around sites critical for nucleic acid binding.

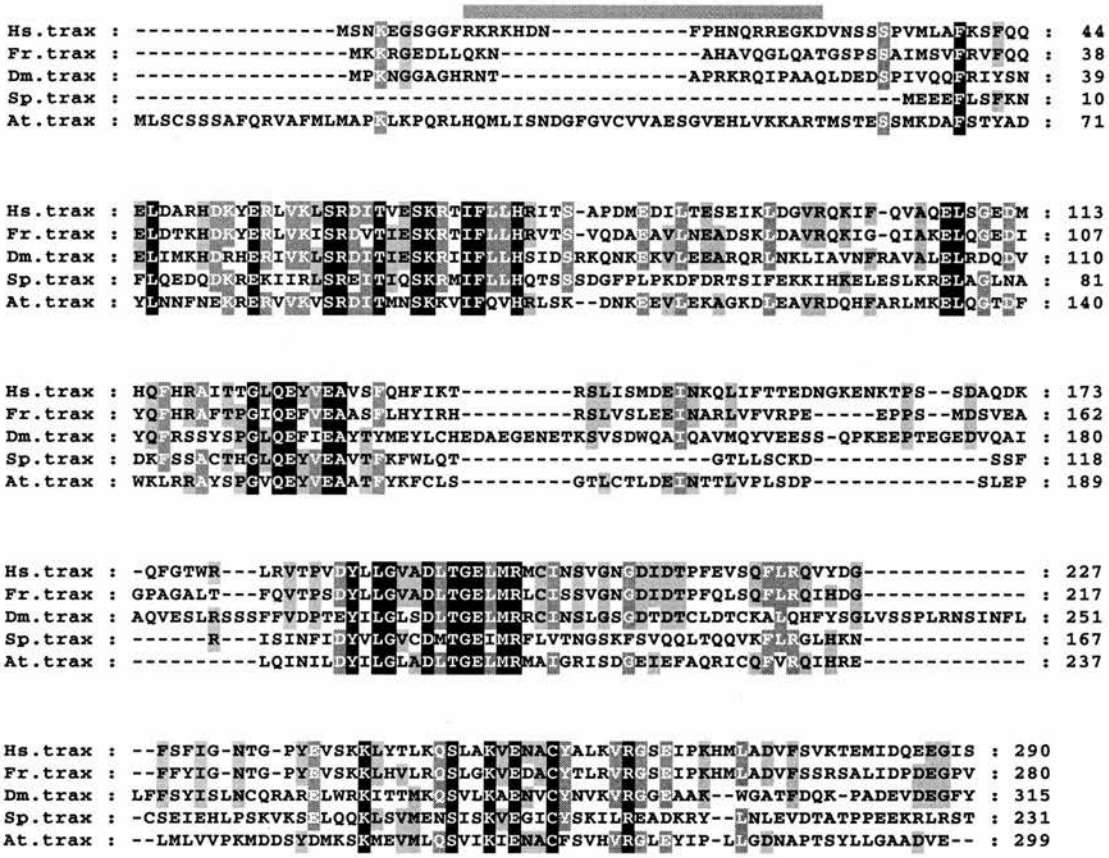


Figure 8.10; Multiple sequence alignment of TRAX orthologues. Numbers to the right of the alignment show the amino acid coordinate of the last residue in the associated section of alignment. Hs indicates *Homo sapiens* (Q99598); Fr, *Fugu rubripes* (Appendix I); Dm, *Drosophila melanogaster* (Q9VF77); Sp, *Schizosaccharomyces pombe* (O74955); At, *Arabidopsis thaliana* (Q9SJK5). Black background indicates 100% amino acid identity, dark grey 80% identity and light grey 60% identity. The grey bar above the alignment indicates the extent of the predicted bipartite nuclear localisation signal in human TRAX. The full length of each protein is aligned. Non-human mammalian TRAX orthologues were excluded from the alignment as their high identity with human TRAX was not informative.

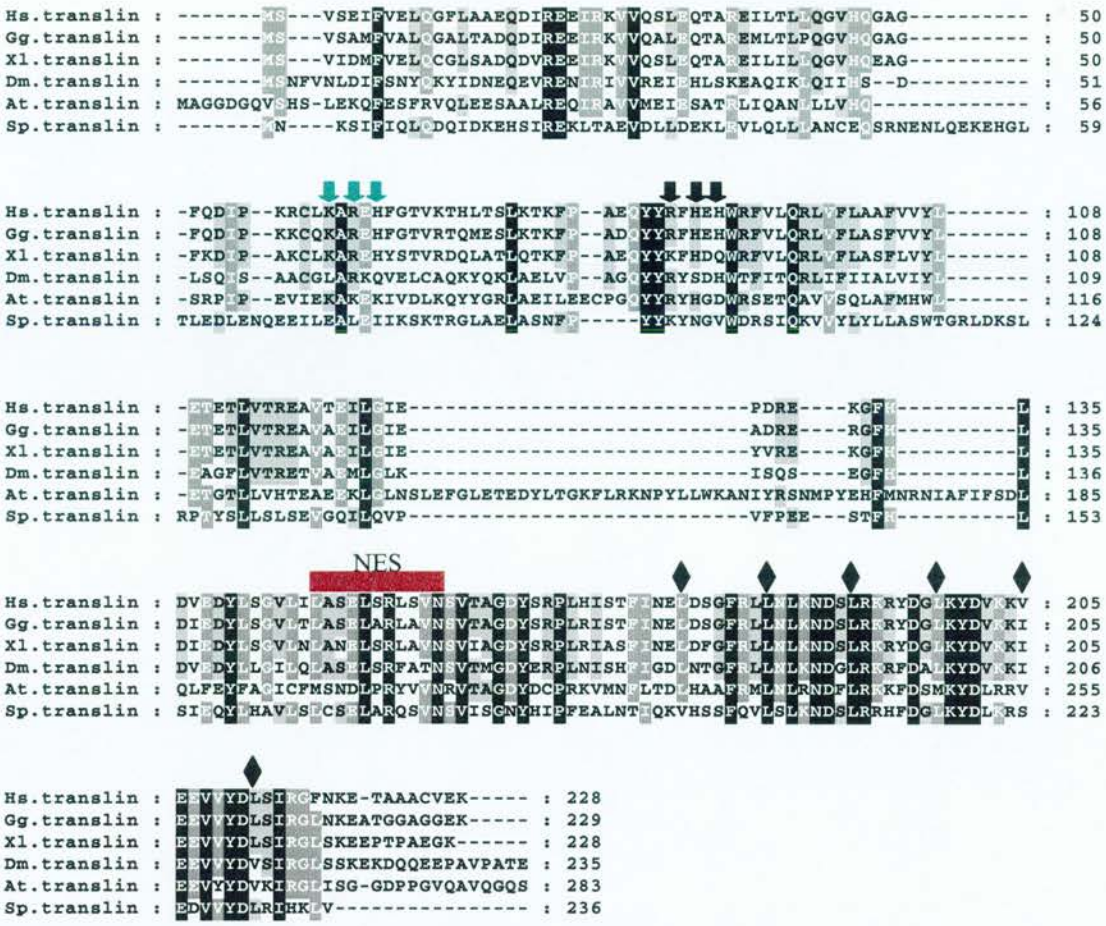


Figure 8.11; Multiple sequence alignment of Translin orthologues. Numbers to the right of the alignment show the amino acid coordinate of the last residue in the associated section of alignment. Hs indicates *Homo sapiens* (Q15631); Gg, *Gallus gallus* (P79769); Xl, *Xenopus laevis* (Q9IAM5); Dm, *Drosophila melanogaster* (Q9V5M0); At, *Arabidopsis thaliana* (Q9ZPQ4); Sp, *Schizosaccharomyces pombe* (Q9P7V3). Black background indicates 100% amino acid identity, dark grey 80% identity and light grey 60% identity. The full length of each protein is aligned. Non-human mammalian Translin orthologues were excluded from the alignment as their high identity with human Translin is not informative. Above the alignment, black arrows indicate the three basic residues that when mutated in human Translin abolish DNA and RNA binding activity (Aoki *et al.*, 1999). Green arrows indicate the three basic residues that when mutated in human Translin abolish DNA binding activity but not RNA binding activity (Chennathukuzhi et al, 2001). The brown bar annotated as NES defines the extent of a functional nuclear export signal in human Translin (Chennathukuzhi et al, 2001). Black diamonds indicate the heptad repeat spacing of leucine (and valine) residues in human Translin that are speculated to form a leucine zipper structure (Aoki *et al.*, 1999).

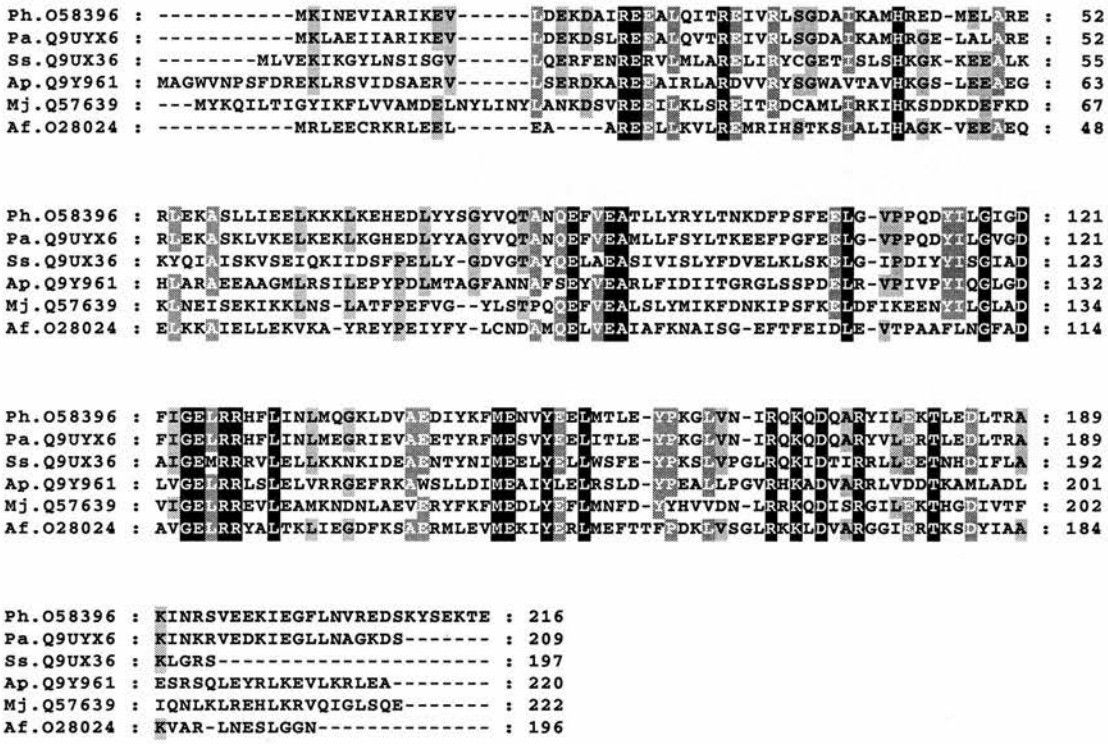


Figure 8.12; Multiple sequence alignment of *Archaea* bacterial Translin / TRAX homologues. Numbers to the right of the alignment show the amino acid coordinate of the last residue in the associated section of alignment. Sequences are referred to by a two character genus and species abbreviation followed by the SPTR accession number of the sequence. Genus and species abbreviations are: Ph, *Pyrococcus horikoshii*; Pa, *Pyrococcus abyssi*; Ss, *Sulfolobus solfataricus*; Ap, *Aeropyrum pernix*; Mj, *Methanococcus jannaschii* and Af, *Archaeaoglobus fulgidus*. *Aeropyrum pernix* and *Sulfolobus solfataricus* are of the *Crenarchaeaota* group of *Archaea* bacteria, the other species shown are of the *Euryarchaeaota* group. Black background indicates 100% amino acid identity, dark grey 80% identity and light grey 60% identity. The full length of each protein is aligned.

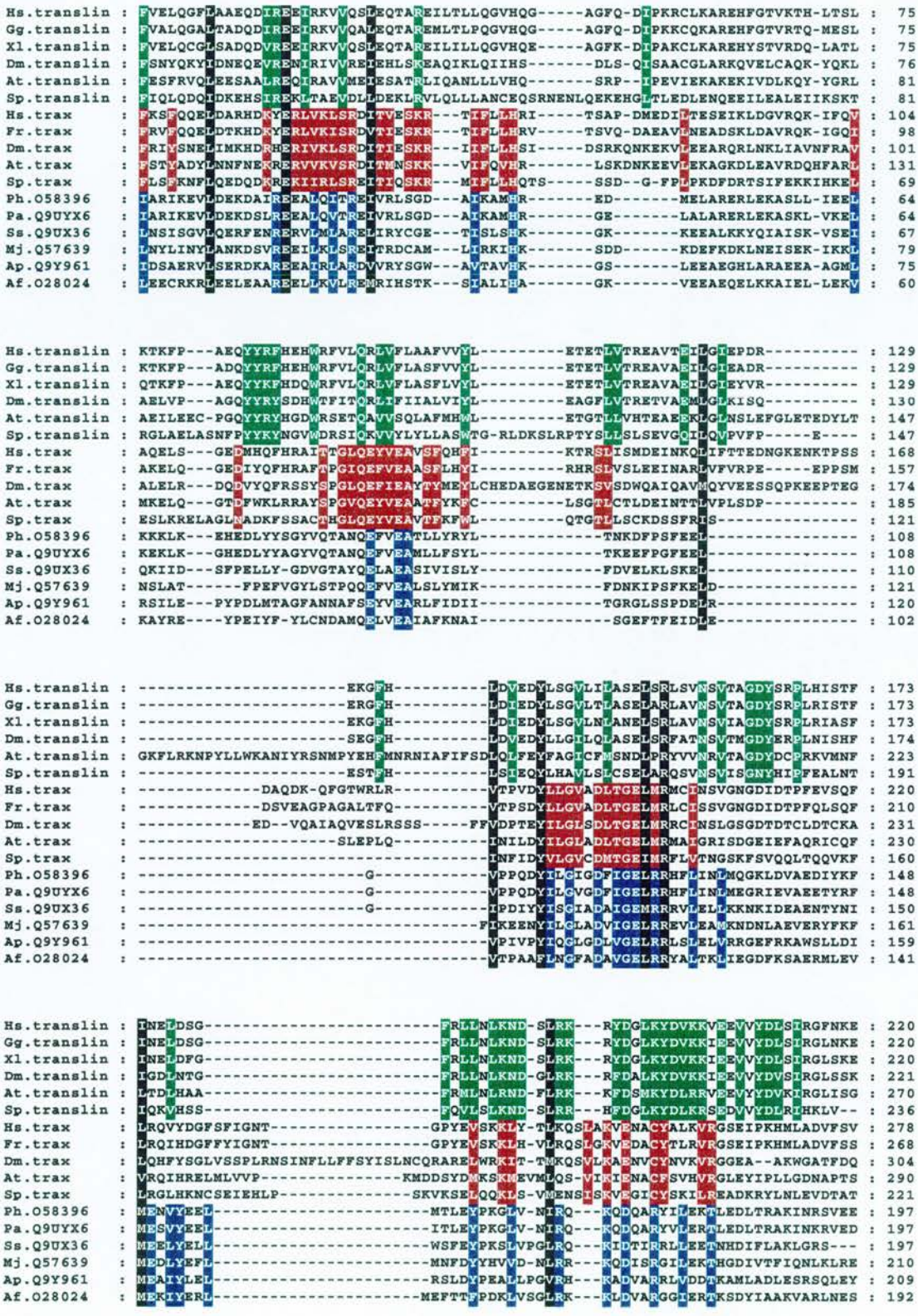


Figure 8.13; Multiple sequence alignment of the TRAX – Translin family of proteins. Full legend on next page.

Figure 8.13; Multiple sequence alignment of the TRAX – Translin family of proteins. Within the alignment, the proteins are clustered into TRAX, Translin and Archaea sub-families. Amino acid similarity is indicated both within and between sub-families. Residues conserved between all aligned sequences are shaded with black background, those conserved within but not between sub-families are shaded with a sub-family specific background: green for Translin, red for TRAX and blue for Archaea. The alignment was created with ClustalW and curated by hand to ensure that within sub-family alignments were not disrupted by between sub-family alignments. Numbers to the right of the alignment show the amino acid coordinate of the last residue in the associated section of alignment. For the proteins annotated as TRAX: Hs indicates *Homo sapiens* (Q99598); Fr, *Fugu rubripes* (Appendix I); Dm, *Drosophila melanogaster* (Q9VF77); Sp, *Schizosaccharomyces pombe* (O74955); At, *Arabidopsis thaliana* (Q9SJK5). For the proteins annotated as Translin: Hs indicates *Homo sapiens* (Q15631); Gg, *Gallus gallus* (P79769); Xl, *Xenopus laevis* (Q9IAM5); Dm, *Drosophila melanogaster* (Q9V5M0); At, *Arabidopsis thaliana* (Q9ZPQ4); Sp, *Schizosaccharomyces pombe* (Q9P7V3). Archaea bacterial proteins are referred to by a two character genus and species abbreviation followed by the SPTR accession number of the sequence. Genus and species abbreviations are: Ph, *Pyrococcus horikoshii*; Pa, *Pyrococcus abyssi*; Ss, *Sulfolobus solfataricus*; Ap, *Aeropyrum pernix*; Mj, *Methanococcus jannaschii* and Af, *Archaeaoglobus fulgidus*.

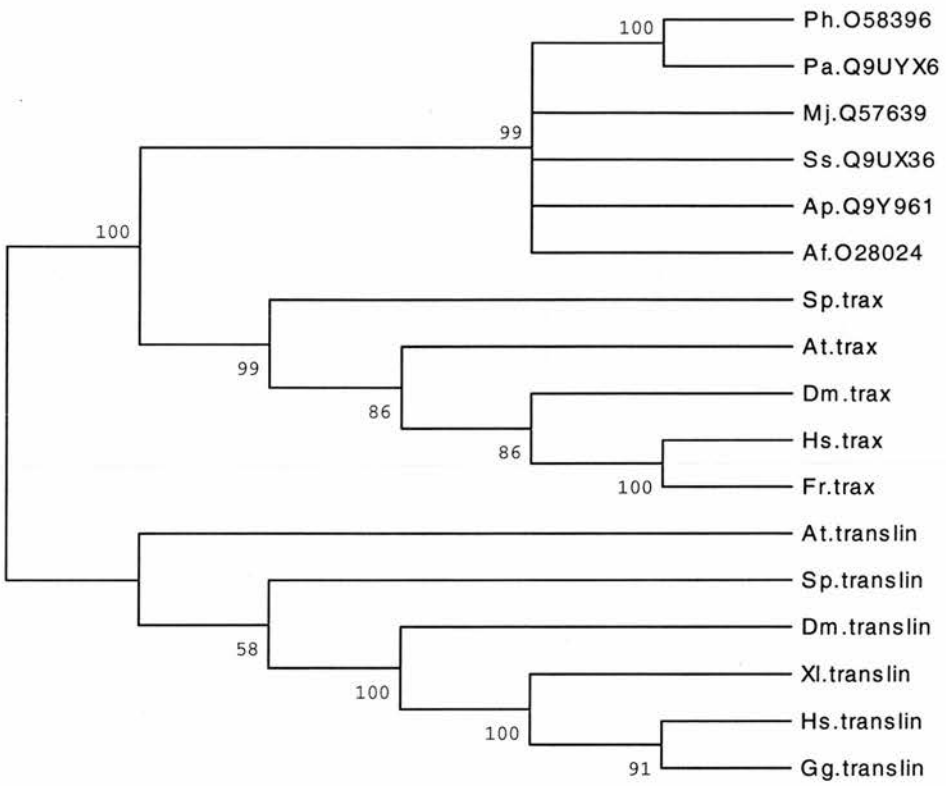


Figure 8.14; Phylogenetic analysis of Translin, TRAX and the homologous *Archaea* bacterial proteins. A neighbour joining, bootstrap consensus tree based on the p-distance method. Numbers at branch forks indicate the percentage bootstrapping support out of 1000 replicates. For the proteins annotated as TRAX: Hs indicates *Homo sapiens* (Q99598); Fr, *Fugu rubripes* (section 8.4.2); Dm, *Drosophila melanogaster* (Q9VF77); Sp, *Schizosaccharomyces pombe* (O74955); At, *Arabidopsis thaliana* (Q9SJK5). For the proteins annotated as Translin: Hs indicates *Homo sapiens* (Q15631); Gg, *Gallus gallus* (P79769); Xl, *Xenopus laevis* (Q9IAM5); Dm, *Drosophila melanogaster* (Q9V5M0); At, *Arabidopsis thaliana* (Q9ZPQ4); Sp, *Schizosaccharomyces pombe* (Q9P7V3). *Archaea* bacterial proteins are referred to by a two character genus and species abbreviation followed by the SPTR accession number of the sequence. Genus and species abbreviations are: Ph, *Pyrococcus horikoshii*; Pa, *Pyrococcus abyssi*; Ss, *Sulfolobus solfataricus*; Ap, *Aeropyrum pernix*; Mj, *Methanococcus jannaschii* and Af, *Archaeaoglobus fulgidus*.

8.6.5 Sequence similarity to nucleic acid helicases

It has previously been speculated that Translin has nucleic acid helicase activity (Aoki *et al.,* 1995). It is interesting then to note that the Translin domain containing proteins of *Archaea* bacteria have significant (E-value 0.0006 when searching with BLASTP, using STPR:Q57639 against the SPTR database) sequence similarity with a group of *Archaea* bacterial nucleic acid helicase proteins (O58731, Q9V019 and Q58969). Each of these helicase proteins contain the helicase-C Pfam domain (PFAM: PF00271) of which there are over 1,369 members known in the public sequence database. Sequence similarity between the *Archaea* Translin domain containing proteins and the helicase proteins extends over the C-terminal half of the helicase-C domain as it is defined in Pfam (figure 8.15, section 2.11.1). The identification of sequence similarity between proteins that are clearly homologous to Translin and share sequence similarity with a family of nucleic acid helicases is particularly noteworthy, as a structural and functional similarity with nucleic acid helicases is consistent with the observation of nucleic acid binding by Translin and its homologous protein TRAX.

To investigate this potential relationship between the Translin domain and nucleic acid helicases further, an alignment was made between the *Archaea* Translin domain sequences and the Translin-like region of the helicase proteins (figure 8.15). Residues were conserved along the length of the alignment, with aliphatic residues being particularly conserved (figure 8.15).

It is functionally plausible that there is an underlying protein structure common to the Translin domain and the Translin-like region of the helicase proteins. Sequence similarity is readily detected (BLASTP) between proteins with the Translin domain and the Translin-like region of the helicase proteins, although the 20 to 35% sequence identity observed in global alignment (figure 8.15) is at the edge of the "grey area" in inferring structural similarity from sequence similarity (Park *et al.,* 1998). Secondary structure prediction of each of the Translin domain sub-families was found to correlate well (figure 8.16), whereas the Translin-like region of the helicase proteins did not (figure 8.16). In the absence of a resolved structure for an

example of the Translin domain it cannot be reliably concluded as to whether the Translin domain is structurally and by inference, functionally related to nucleic acid helicases.

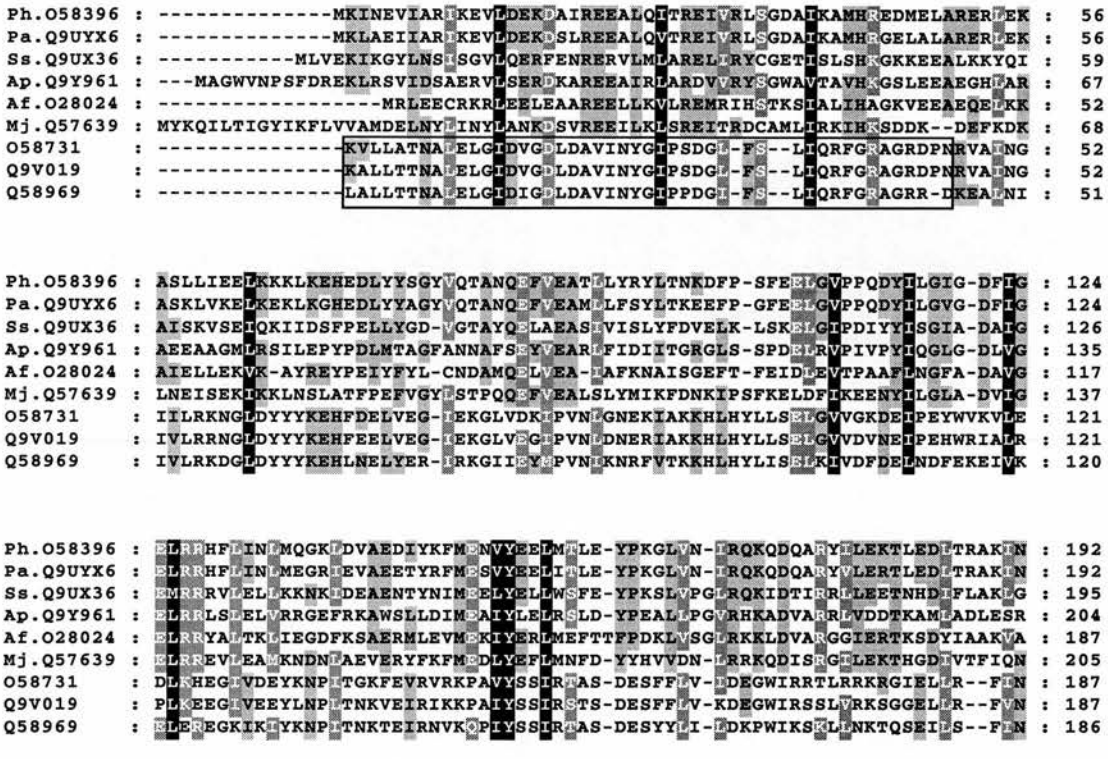


Figure 8.15; Alignment of *Archaea* Translin domain proteins and the Translin-like region of the nucleic acid helicase proteins. Black background indicates 100% conservation, dark grey 80% and light grey 60%. The region of the nucleic acid helicase proteins that overlaps the helicase-C Pfam domain (Pfam: PF00271) is indicated by a frame over the sequence. Translin domain sequences are referred to by a two character genus and species abbreviation followed by the SPTR accession number of the sequence. Genus and species abbreviations are: Ph, *Pyrococcus horikoshii*; Pa, *Pyrococcus abyssi*; Ss, *Sulfolobus solfataricus*; Ap, *Aeropyrum pernix*; Mj, *Methanococcus jannaschii* and Af, *Archaeaoglobus fulgidus*. The helicase protein sequences are indicated by the SPTR accession number. Only regions corresponding to amino acids 331 to 532 of O58731, 330 to 530 of Q58969 and 330 to 531 of Q9V019 are shown in the alignment. The complete sequence of Translin domain proteins is shown.

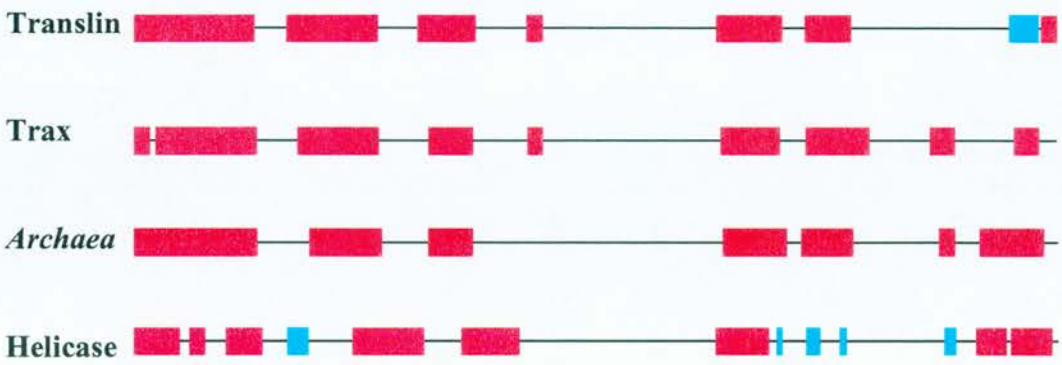


Figure 8.16; Aligned secondary structure prediction of Translin domains and helicase proteins. PHD (section 2.11.2) predictions of secondary structure based on the alignments shown in figures 8.10, 8.11 and 8.12 of Translin, TRAX and *Archaea* protein sequences. Helicase prediction is based on clustalw (section 2.11.2) alignment of amino acid sequences SPTR: O58731, Q9V019 and Q58969.Red boxes indicate predicted helix secondary structure. Blue boxes indicate predicted β-sheet secondary structure.

8.7 Discussion

Work presented in this chapter has identified and characterised the *TRAX* gene in mouse, *Fugu* and *Tetraodon,* as well as identifying homologous sequences in a wide range of eukaryotes based on sequence similarity searching in public sequence databases. Expression of the *TRAX* gene was demonstrated in *Fugu* and the genomic structure resolved, revealing a potential example of intron sliding in the *Fugu TRAX* gene. Intron 1 of *TRAX* has an exceptionally consistent size between humans, *Fugu* and *Tetraodon* and shows sequence similarity along its entire length between these species. The largest of *TRAX* introns, intron 4 was also found to harbour non-coding conserved sequences. A small conserved sequence motif was also found in the 3' UTR of human, mouse, *Fugu* and *Tetraodon* transcripts. It is likely that introns 1 and 4 conserved elements are regulators of *TRAX* transcription, the 3' UTR element has the potential to regulate expression at the transcriptional and post transcriptional level.

Several intergenic conserved sequences were identified, particularly in the *TRAX – DISC1* intergenic region. These conserved intergenic sequences are good candidates for transcriptional regulators of the neighbouring genes: *EGLN1*, *TRAX* and *DISC1*.

There was a specific interest in the potential conservation of intergenic splicing between *TRAX* and *DISC1*. Sequences representing the human and mouse intergenic exons were not conserved in *Fugu*. Specifically testing for the intergenic splicing of *TRAX* to *DISC1* in *Fugu* failed to detect any such transcripts, while *TRAX* and *DISC1* transcripts could both be detected in the tissues tested. The absence of sequence conservation between human intergenic exons and *Fugu* genomic sequence, lack of a continuous reading frame and inability to detect intergenic splicing in *Fugu*, demonstrate that the intergenic splicing of *TRAX* and *DISC1* is not a feature that has been well conserved during the evolution of this locus.

From comparative genomic alignment, there is no evidence for protein coding genes between *EGLN1* and *DISC1*, other than the previously identified *TRAX* gene. The

Backtrax and *Foretrax* transcripts identified through human EST alignment to genomic sequence are not conserved and are unlikely to be protein coding transcripts. Their functional significance remains unclear.

The TRAX protein was demonstrated to be a member of the Translin domain family of proteins, with novel homologues identified in *Archaea* bacteria. Sequence similarity was also detected between the Translin domain and a large family of nucleic acid helicases. The level of sequence similarity and consistency of secondary structure prediction between the helicases and the Translin domain was insufficient to conclude that the sequence similarity reflected homology. However, the known nucleic acid binding properties of TRAX and Translin are consistent with helicase like properties. The issue of whether the Translin domain represents a helicase structure is only likely to be resolved if new sequences that represent intermediates between the helicases and the Translin domain are identified, or if the crystal structure of a Translin domain is obtained.

The role of TRAX in binding ssDNA in the context of double stranded DNA breaks is not a function that is readily reconciled with an end phenotype of major mental illness. Translin not only binds ssDNA, but also mediates the transport, sub-cellular localisation and translational masking of specific mRNA species within neurons and other tissues. The dendritic localisation of specific mRNAs has been proposed to be fundamental for the structural and functional polarity of neurons (Kuhl and Skehel, 1998 for review). The translational masking of transcripts combined with their anchoring to the cytoskeleton within specific regions of a cell (particularly neural dendrites) are all properties of Translin and represent the hypothetical properties of a mechanism for mediating synaptic plasticity (Kuhla and Skehel, 1998).

Although TRAX is not able to bind mRNA and mediate the translational masking and sub-cellular transport like its paralogue Translin, it is able to modify the behaviour of Translin such that it can no longer bind RNA and may even mediate its translocation from the cytoplasm to the nucleus (section 8.2.3). Through this means, TRAX could indirectly influence dendritic mRNA localisation. For these reasons,

TRAX and Translin can be considered functional candidates for cognitive and psychiatric illness. By the same rationale other TRAX / Translin interacting proteins could also be considered functional candidates, particularly products of the *PURA* gene which may be antagonistic to the action of Translin homo-dimers in RNA binding (Ohashi *et al.,* 2000).

Translin is located on chromosome 2q21 (Aoki *et al.,* 1997) and *PURA* on 5q31 (Ma *et al.,* 1995). There is no strong evidence for a susceptibility locus for major mental illness around the Translin gene on 2q21, although two reports have found suggestive evidence for a susceptibility gene in the vicinity of 2q21 (Aschauer *et al.,* 1993; Levinson *et al.,* 1998). Also a balanced translocation t(2;18) that co-segregates with schizophrenia in a small family has been reported (Genest *et al.,* 1976) with the chromosome 2 breakpoint cytogenetically mapped to 2q21. In contrast, the 5q31 locus of *PURA* has repeatedly been implicated as the site of a susceptibility gene for major mental illness (Straub *et al.,* 1997; Eckstein *et al.,* 1997; Schwab *et al.,* 1997; Edenberg *et al.,* 1997; Hovatta *et al.,* 1998; Morisset *et al.,* 1999; Crowe and Vieland, 1999 for review). In particular, the maximum LOD score of 3.13 for schizophrenia in 265 Irish pedigrees (Straub *et al.,* 1997) and an NPL (non-parametric LOD score) of 3.76 for bipolar affective disorder in six pedigrees (Byerley *et al.,* personal communication reported in Crowe and Vieland, 1999), provide strong evidence for a major mental illness susceptibility locus in the region of chromosome 5q31.

The functional rationale for an involvement of TRAX, Translin and associated proteins in the molecular aetiology of mental illness would appear to be at least as good as other functional candidates that have been considered (section 1.5.4). The coincidence of *TRAX* and *PURA* in regions of the genome strongly implicated as susceptibility loci warrants their consideration as positional and functional candidates worthy of further genetic investigation.

Chapter 9

The EGLN gene family

9.1 Preface

A novel homologue of *C. elegans* Egg laying-9 (*Egl-9*) was identified in *Fugu* genomic sequence directly upstream of *TRAX* and *DISC1* (section 4.2). Subsequently a close homologue of this gene, provisionally named *C1orf12* (Dupuy *et al.,* 2000) and subsequently referred to as *EGLN1*, was found in human genomic sequence upstream of the *TRAX* gene (section 4.2 and Dupuy *et al.,* 2000). Given the proximity of *EGLN1* to the chromosome 1 translocation breakpoint it was considered a positional candidate along with *DISC1*, *DISC2* and *TRAX* for involvement in the psychiatric phenotype of t(1;11) translocation carriers.

Preliminary searching of EST and genomic sequence databases identified multiple homologous sequences that had not been characterised or annotated. An investigation of these *Egl-9* homologous sequences was carried out in order to evaluate *EGLN1* as a candidate for the t(1;11) phenotype and understand the biological roles of the genes at this locus

9.2 Introduction

Egl-9 was originally isolated as the gene responsible for an egg laying defective phenotype of the nematode worm *C. elegans* (Trent *et al.,* 1983). Identified along with 58 other characterised loci (*Egl-1* to *Egl-59*), *Egl-9* mutants are fertile, but in the hermaphrodite worms release of progeny is delayed compared to wild type.

More recently Darby *et al.,* (1999) have shown that at high cell density some strains of *Pseudomonas aeruginosa* and *Pseuodmonas fluorescens* produce a toxin that efficiently kills *C. elegans*. Loss of function mutations in Egl-9 confer a strong

resistance to this toxicity (Darby *et al.,* 1999). When wild type *C. elegans* were placed on a lawn of *P. aeruginosa* (strain PA01), the normally rhythmical pharyngeal pumping became sporadic within seconds. A progressive paralysis leads to terminal paralysis within four hours. Separating the worms from *P. aeruginosa* with 0.025 μm pore nitrocellulose filters failed to prevent the lethal neuro-muscular paralysis, demonstrating that the toxin is diffusible (Darby *et al.,* 1999). This evidence points to *Egl-9* being the target or a mediator of an as yet unidentified diffusible toxin.

SM-20, a vertebrate homologue of *Egl-9*, has been implicated in the differentiation and growth regulation of muscle cells in the rat (Wax *et al.,* 1994 and Moschella *et al.,* 1999) and has subsequently been shown to be necessary for nerve growth factor dependent survival of neurons (Lipsomb *et al.,* 1999). However, over expression can induce apoptotic cell death in neurons (Lipscomb *et al.,* 1999). *SM-20* appears to be a downstream mediator of p53 signalling. *SM-20* transcription is dramatically up-regulated by the activation of temperature sensitive p53 and over expression of rat *SM-20* in cells lacking functional p53 induces cell growth suppression (Madden *et al.,* 1996). *SM-20* induced cell death has been shown to be accompanied by caspase-3 activation and inhibition of caspase activity prevents *SM-20* induction of apoptosis (Lipscomb *et al.,* 2001) implying that *SM-20* acts between p53 and caspase 3 in a signalling pathway.

Further investigation of this neural, apoptotic induction by Lipscomb *et al.,* (2001) has led to a convincing demonstration of mitochondrial targeting of SM-20. The mitochondrial targeting signal of SM-20 has been mapped by fusion and deletion experiments to the first 25 amino acids of the protein (Lipscomb *et al.,* 2001). This 25 amino acid region was observed to be rich in hydroxylated and basic amino acids and devoid of acidic amino acids, as is typical for mitochondrial targeting sequences (Hurt & Schatz 1987). It was also shown by deletion of the mitochondrial targeting sequence that mitochondrial localisation is not necessary for the role of SM-20 in apoptotic induction (Lipscomb *et al.,* 2001). In the absence of a mitochondrial targeting signal, SM-20 was observed to localise to both the cytosol and nucleus. A

model for the role of SM-20 in the induction of neuronal apoptosis has been proposed by R. Freeman (personal communication) and is outlined in figure 9.1.

C. elegans egl-9 (O45918), rat *Sm-20* (AAG33965) and human *SCAND2* genes have previously been described (Trent *et al.,* 1983; Lipscomb *et al.,* 1999; Dupuy *et al.,* 2000). It is however apparent from BLAST (Altschul *et al.,* 1990) homology searching, that there are multiple other homologous sequences represented in the EST and genomic sequence databases. This study, set out to identify and characterise the human and mouse homologues and investigate the evolutionary history of this interesting and little understood gene family.

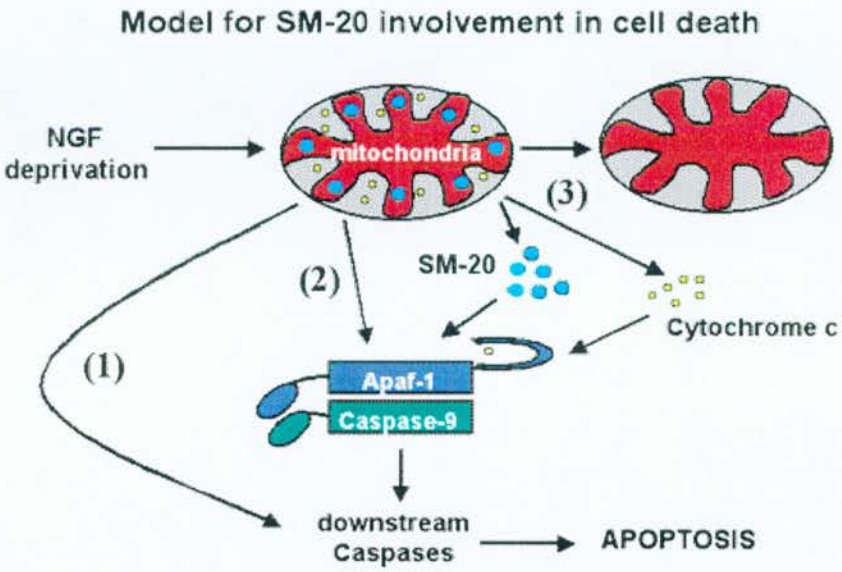


Figure 9.1; Model for SM-20 involvement in neuronal cell death. Reproduced with permission, R. Freeman (University of Rochester, New York US). Expression of *SM-20* is upregulated during neuronal apoptosis and its overexpression in neurons causes cell death. SM-20 is a mitochondrial protein that induces cell death through activation of cysteine proteases called caspases. SM-20 could function as an indirect activator of downstream **(1)** or upstream **(2)** caspases. Alternatively, it may be released from mitochondria into the cytosol during cell death **(3)** to function as a caspase activator.

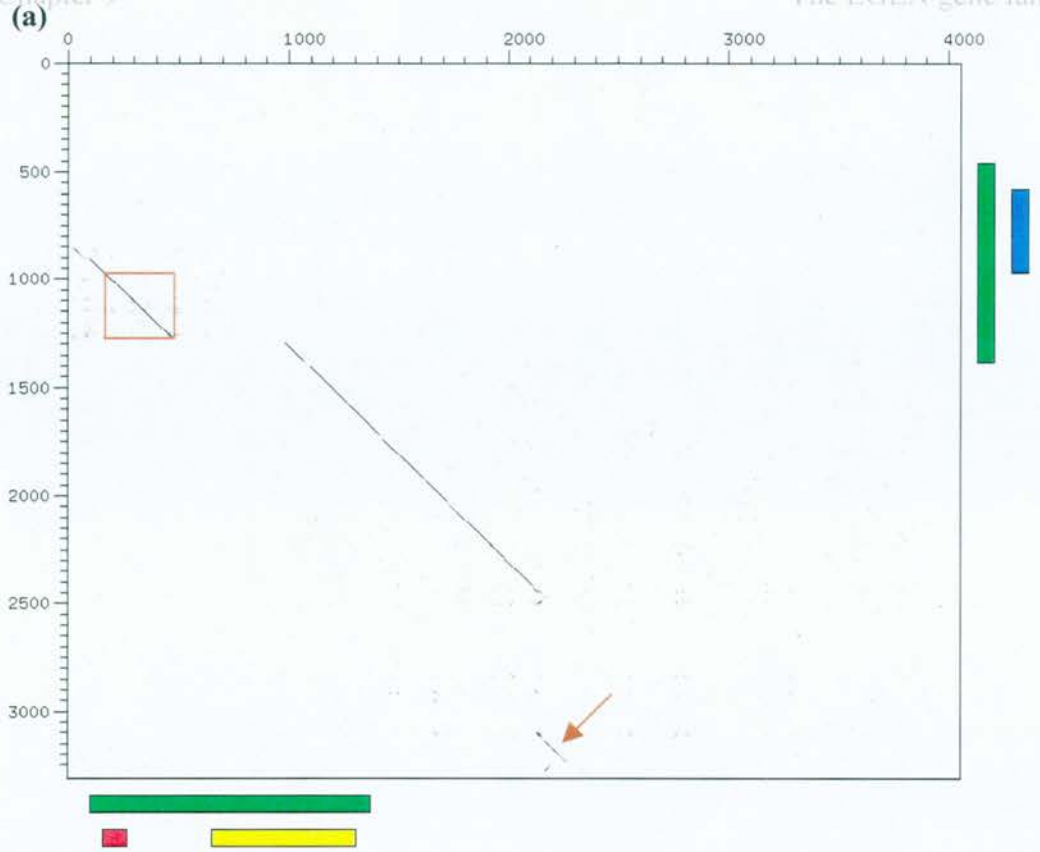
9.3 Identification of human and mouse *Egl-9* homologues

9.3.1 EST clustering

The clustering of ESTs was seeded by an initial search using *C. elegans* Egl-9 and rat SM-20 amino acid sequences as TBLASTN (section 2.11.2) queries against mouse or human sub-sets of the EMBL EST data set (section 2.11.1). All EST matches with a TBLASTN bit-score of 58 or greater were used to initiate the sequence clustering process described in section 2.11.4. A TBLASTN bit-score cut off of 58 was used as it represented a boundary in the distribution of scores between the best matches (including all known homologues) and all other sequences when Egl-9 was searched against human ESTs.

The EST assemblies were manually curated to ensure that closely related genes were not assembled into the same contig, this was essential to separate *SCAND2* from *C1orf12* transcripts which were >90% identical in local alignment (figure 9.2) and had several stretches of 100% identity for >80 nucleotides.

Assemblies representing four human genes and three mouse genes were generated from the clustered ESTs. One of the human assemblies which was comprised of eleven ESTs and four mRNA database entries represents the previously described *SCAND2* transcript (Dupuy *et al.,* 2000), which although derived in part from a human homologue of *Egl-9*, is likely to be translated in an altered reading frame (Dupuy *et al.,* 2000) and consequently would not produce an Egl-9 related protein product (section 9.2.6).



(b)

```
        110        120        130        140        150        160
EGLN1   CCGCAGCTCCTTCTACTGCTGCAAGGAGCACCAGCGTCAGGACTGGAAGAAGCACAAGCT
        ::::::::::::   :::::::::::::: :::::: ::::::::::::::::::::::::
SCAND2  CCGCAGCTCCTT---CTGCTGCAAGGAGCGCCAGCGCCAGGACTGGAAGAAGCACAAGCT
         1010          1020       1030       1040       1050       1060

        170        180        190        200        210        220
EGLN1   CGTGTGCCAGGGCAGCGAGGGCGCCCTCGGCCACGGAGTGGGCCCACACCAGCATTCCGG
        :::::::::::::::::::::::::::::::::::::::: :::::: :::::: : :::::
SCAND2  CGTGTGCCAGGGCAGCGAGGGCGCCCTCGGCCACGGAGGGGGCCCTCACCAGGACTCCGG
             1070       1080       1090       1100       1110       1120

        230        240        250        260        270        280
EGLN1   CCCCGCGCCGCCGGCTGCAGTGCCGCCGCCCAGGGCCGGGGCCCGGGAGCCCAGGAAGGC
        :::::::::::: ::::::: ::::::: :::::: : :::::: :::: ::::::::::
SCAND2  CCCCGCGCCGCCCGCTGCAGCGCCGCCGTCCAGGGACCGGGCCCTGGAGGCCAGGAAGGC
             1130       1140       1150       1160       1170       1180

        290        300        310        320        330        340
ELGN1   AGCGGCGCGCCGGGACAACGCCTCCGGGGACGCGGCCAAGGGAAAAGTAAAGGCCAAGCC
        ::::  :::::::::::: :::::::::::::::: :::::::      : :::::::::: :
SCAND2  AGCGAGGCGCCGGGACAGCGCCTCCGGGGACGCAGCCAAG------GCAAAGGCCAAGTC
             1190       1200       1210       1220            1230
```

Figure 9.2; High degree of sequence identity between *C1orf12* and *SCAND2*. **(a)** Dotter alignment of the complete human *EGLN1* and *SCAND2* transcripts. **(b)** Local alignment of a region conserved between *EGLN1* and *SCAND2*. See over leaf for full legend.

Figure 9.2; Sequence identity between *EGLN1* and *SCAND2*. **(a)** Dotter alignment of the complete human *EGLN1* transcript (AJ310543) (horizontal) and the *SCAND2* transcript (AF229246) (vertical). The scale on both axes indicate nucleotide position (nt) along the transcripts. The major open reading frame of each transcript is indicated by the green rectangle. The conserved EGLN domain (section 9.5.3) is marked by the yellow rectangle, the MYND-type zinc finger by the red rectangle and the SCAND domain by the blue rectangle. The region of alignment indicated by the brown box indicates sequence shown in panel 'b'. The arrow indicates an alignment of Alu repetitive elements. **(b)** Local alignment (Fasta33, section 2.11.2) of human *EGLN1* and *SCAND2*. The region of alignment shown corresponds to the sequence indicated in panel 'a', which is predicted to be coding sequence in both transcripts but translated in differing reading frames (section 9.3.6).

9.3.2 Genomic sequence clustering

Preliminary examination, based on sequence associated annotation, of *Egl-9* homology in human genomic sequence suggested that *Egl-9* homologues are located at more than nine discreet chromosomal locations. *C. elegans* Egl-9 and rat SM-20 amino acid sequences were used as queries for TBLASTN (section 2.11.2) searching of all publicly available human genomic sequence (from HTG, GSS and EMBLminus subsets of the EMBL database; section 2.11.1). The 17 genomic clones identified (table 9.1) were clustered based on FPC and sequence overlaps as determined in the Accession Maps project (http://genome.wustl.edu/gsc/human/) based on the October 7th 2000 data freeze. This clustering suggested that there are five rather than nine regions in the human genome with strong homology to Egl-9 and SM-20, on chromosomes 1, 12, 14, 15 and 19.

9.3.3 Integrating genomic and cDNA data

Of the four human EST assemblies (table 9.2), one comprised of eleven ESTs and four mRNA sequences represents the previously described *SCAND2* transcript (Dupuy *et al.,* 2000). Nucleotide identity between pairs of human and mouse EST assemblies, strongly suggests that the three remaining paralogous human transcripts (excluding SCAND2) are orthologous to the three mouse transcripts. These genes are hereafter termed Egg Laying Nine-1, Egg Laying Nine-2 and Egg Laying Nine-3 (*EGLN1*, *EGLN2* and *EGLN3* respectively) (HUGO accepted nomenclature) with reference to their mutual homology with *Egl-9*.

The human transcript for *EGLN1* aligns (Sim4, section 2.11.2) to genomic sequence from human chromosome 1 identified in section 9.2.2 (table 9.1) with >99% identity, gaps in the alignment are consistent with canonical splice site consensus sequences. Similarly, *EGLN2* aligns to chromosome 19, *EGLN3* to chromosome 14 and *SCAND2* to chromosome 15 with sufficient identity (>99%) to determine that each of these genomic loci represent the template for respective transcripts. The remaining region of the human genome with homology to *Egl-9* and *SM-20*, was from chromosome 12 (table 9.1). This sequence was not represented by ESTs and likely represents a processed pseudogene (section 9.3.8).

Gene[a]	Genomic sequences[b]	Chromosome annotation[c]	Accession Map annotation[d]		
			Contig[e]	Chromosome[f]	RH map[g]
EGLN1	AL445524	1	NA	-	-
	AL358784	1	ctg13079	1	736.71
	AL117352	1q42	ctg13079	1	736.73
	AC011945	NA	ctg13079	1	736.84
	AC012242	NA	ctg13079	-	-
EGLN2	AC008537	19	ctg15547	19	235.50
	AC025769	5	ctg14739	19	236.31
	AC019337	NA	ctg17685	19	235.44
EGLN3	AC022969	14	ctg53	14	65.90
	AL358340	14	ctg53	14	65.90
	AC023450	1	ctg53	-	-
	AC084333	4	NA	-	-
SCAND2	AC048382	18	ctg12246	15	300.12
	AC016771	15	ctg12246	15	300.43
	AC087732	15	NA	-	-
ψEGLN3	AC018654	12	ctg14210	12	69.39
	AC022073	12	ctg14210	12	66.75

Table 9.1; Genomic content and location of the human EGLN gene family. **(a)** Gene name. **(b)** Sequence accession numbers for each genomic clone with significant (section 2.11.4) homology to *C. elegans* Egl-9 and rat SM-20, clustered by sequence identity. **(c)** Map annotation for each of the sequence accessions. **(d)** Data derived from the Accession Maps project (http://genome.wustl.edu/gsc/human/Mapping/). **(e)** Accession Map contig of the sequence, based on the October 7th 2000 data freeze. **(f)** Chromosomal location as determined by STS content of the clone sequence. **(g)** Radiation hybrid (RH) map location of the clone.

Gene[a]	mRNA[b]	EST tiling path[c]
Human *EGLN1*	AF229245 AF334711 AF277176 AJ227859 AF277174	AI343586, AF229245, BE669740, BE960611, AA508346, BG117305, AA370652, AA203627, AW377161, AA218859
Human *EGLN2*	AK026863 AK025396	BF061631, BE561402, AW957364, AA312497, AI110596, AL40033
Human *EGLN3*	AK025273 AX035283 AK026918	BF726360, BF724589, R00332, AW079532, A804037
Mouse *EGLN1*		BF465779, AW610745, AA434738, BF022323, BF161755, BF782795
Mouse *EGLN2*		BF321818, BF301163, AW320216, AA939949, AA014327, BF730642
Mouse *EGLN3*		BF540183, BF142691, W14123, AI020522

Table 9.2; Minimal tiling path of EST assemblies. **(a)** Mouse and human *Egl-9* related genes identified by EST clustering. **(b)** cDNA sequences representing the transcription of these genes. **(c)** Tiling path of ESTs from each assembly, from which the complete assemblies can be derived.

9.3.4 *EGLN1*

Human *EGLN1* (*C1orf12*) has been reported elsewhere and its genomic structure described (Dupuy *et al.,* 2000). In northern blot analysis, a major band at ~5 kb and a less abundant ~2.4 kb band was observed (Dupuy *et al.,* 2000). Assembly of 60 human ESTs results in a contiguous assembly of 4029 bp. The addition of a poly A tail and additional 5' UTR not represented in the EST assembly, would correspond to the observed ~5 kb transcript. Four of the ESTs indicated that a second poly-adenylation site was also used. This produced a transcript of 1879 bp, which probably represented the less abundant ~2.4 kb transcript.

Aligning the *EGLN1* EST assembly with the transcript reported by Dupuy *et al.,* resulted in a predicted transcript of 7089 bp, not including poly-adenlyation. This is at least 2 kb longer than the major transcript detected by Northern blot. The compelling evidence from ESTs was for a major transcript that contains all of the assembled 4029 bp of sequence. The transcription initiation site proposed by Dupuy *et al.,* was approximately 2 kb upstream of the major transcriptional start site inferred here.

Mouse *EGLN1* was assembled from a cluster of 97 ESTs, into a 2835 bp contig that is incomplete at the 5' end and demonstrated the use of at least 2 poly-adenylation signals. Both human and mouse *EGLN1* transcripts make use of a non-consensus ATTAAA polyadenylation signal. Within the 3' UTR of the mouse and human transcripts there was a block of approximately 300 nt showing 86 % identity between species (not shown). This was substantially greater conservation than the remainder of the 3' UTR and may reflect a site of regulatory importance.

9.3.5 *EGLN2*

There are 181 ESTs and 2 full insert sequence (FIS) cDNA clones derived from this gene present in the human *EGLN2* cluster. These assemble into a 2086 bp transcript with a 407 codon open reading frame. A single EST (BF690531) represents the skipping of exon 4. This would shift the reading frame and cause exon 5 to be translated in an altered reading frame for 29 codons before encountering a stop codon. The cDNA clone AK026863 retains intron 3, although other introns are spliced out, causing the premature termination of the reading frame 16 codons down stream of the normal exon 3 splice site. The mRNA AL133009 represents an un-spliced transcript from the *EGLN2* locus that initiates in intron 1 and reads through exons 2 to 5. However, the reading frame is not maintained in this transcript, which may represent genomic contamination of the cDNA library.

115 mouse ESTs assembled into a 2089 bp contig representing *EGLN2*. As was found to be the case for the human assembly, there were several ESTs that clearly derived from the *EGLN2* locus that exhibited differential splicing and intron retention. However, each of the alternatively processed transcripts were represented

at less than 1% and none were expected to maintain an open reading frame. Both mouse and human *EGLN2* genes were found to have consensus Kozak motifs (Kozak, 1996) at the predicted translation initiation sites and both utilised a single consensus polyadenylation signal for cleavage and polyadenylation.

9.3.6 *EGLN3*

Human *EGLN3* was identified from a cluster of 55 ESTs, assembled into a contig of 2773 bp. The consensus sequence of the assembly is comprised of 327 bp 5' UTR, a 239 codon open reading frame and 1730 bp of 3' UTR. Mouse *EGLN3* was represented by 49 ESTs, which were assembled into a 2655 bp contig. Both mouse and human assemblies showed that two polyadenylation sites were used. In humans this shows a preference for the proximal site, while in mouse there is a preference for the distal site. In both species the two polyadenylation sites conform to the AATAAA consensus. The predicted translation initiation sites for both mouse and human conform to an adequate context (Kozak, 1996).

9.3.7 *SCAND2*

The *SCAND2* gene was first described by Dupuy *et al.,* 2000 as a novel gene produced through the retrotransposition of *EGLN1* (*C1orf12*) and its subsequent integration downstream of SCAN domain encoding exons. There are 25 known human SCAN domain containing proteins (PFAM version 6.5), with the exception of SCAND2 and an unnamed protein (SPTR: O60388) all known full length SCAN domain proteins also contain C2H2-type zinc fingers (PFAM: PF00096). To date the SCAN domain has only been identified in vertebrates.

Assembled human genomic sequence over the *SCAND2* locus on chromosome 15 (UCSC Golden Path assembly, 12.12.2000) demonstrates that a close human homologue of the mouse *ZFP-29* gene (89% identical and 93% similar in translation) is directly upstream of the *SCAND2* gene. The proteins encoded by both mouse *ZFP-29* and its close homologue (subsequently human *ZFP-29*) contain an N-terminal SCAN domain and 14 tandem C2H2-type zinc finger domains. The SCAN domain is completely contained within the first coding exon of human ZFP-29.

The simplest explanation for the evolution of the *SCAND2* gene is through the duplication of *ZFP-29* exon 1, the reverse transcription and integration of an *EGLN1* transcript downstream of the duplicated exon 1. Within *SCAND2* transcripts, the reading frame of the integrated *EGLN1* transcript is shifted so an EGLN domain containing protein is not encoded, the *EGLN1* reading frame is also disrupted by deletions and point mutations that introduce stop codons. A 306 codon open reading frame, contributed to by the SCAN domain encoding exon 1 and the continuation of that reading frame in the integrated EGLN sequence. It is not known if the *SCAND2* transcript produces a protein or if a *SCAND2* protein has a biological function.

9.3.8 Pseudo-EGLN genes

Sequence from the overlapping BAC clones (embl:AC018654 and embl:AC022073) show 84% identity at the nucleotide level to human *ELGN3* (data not shown). However, a lack of introns, multiple in-frame stop codons, lack of representation in the EST databases and an Alu element insertion into the coding sequence strongly imply that that this is a processed pseudogene of *EGLN3*.

9.4 Gene structure of the EGLN genes

9.4.1 Conserved gene structure

The genomic structures of the human *EGLN* genes were determined through aligning the EST assemblies to draft human genome sequence (Est2genome; Sim4, section 2.11.2), taking into account splice site consensus sequences to resolve alignment ambiguities. *EGLN1* was found to consist of 5 exons, which agrees with that previously reported (Dupuy *et al.,* 2000). *EGLN2* and *EGLN3* genomic structures show the same pattern as *EGLN1*, having a large first exon and phase 0 intron. The second intron is also phase 0, intron 3 is phase 2 and intron 4 is phase 1. In each case, exon 5 has a stretch of coding sequence and subsequently runs directly into 3' UTR without further splicing of the transcript. The conserved position of introns relative to the coding sequence is summarised in figure 9.5.

9.4.2 The gene structure of *Fugu EGLN1*

By homology to human *EGLN1* amino acid sequence, the *Fugu* sequence contig (section 3.5.6) was predicted to contain the first and second exon of *Fugu EGLN1* (figure 9.3). The poor conservation of EGLN1 orthologues between the N-terminal MYND-type zinc finger and C-terminal EGLN domain (figure 9.5) resulted in poor homology based prediction of gene structure (data not shown). Genscan (section 2.11.2) *ab initio* prediction on genomic sequence predicted a first exon that was highly similar to human exon 1 but the sequence with strong similarity in translation to human exon 2 was not predicted as an exon. *A priori* knowledge of human *EGLN1* gene structure was therefore used to predict the boundaries of *Fugu EGLN1* exons 1 and 2.

Sequence corresponding to the predicted exon 1 of *Fugu EGLN1* was used as a hybridisation probe to screen the 11 *Fugu* cDNA libraries detailed in section 2.9.7. No strongly positive clones were detected at high stringency hybridisation (data not shown). In the absence of a cDNA clone to validate the gene structure predictions and demonstrate expression of the *EGLN1* gene, RT-PCR was performed between predicted exons 1 and 2 (figure 9.3). The RT-PCR product was cloned and sequenced demonstrating that this did represent transcription from the *Fugu EGLN1* locus. Alignment of cDNA to genomic sequence demonstrated that the phase and relative position of intron 1 was conserved between human and *Fugu EGLN1*, confirming the predicted, partial gene structure of *Fugu EGLN1*.

The partial open reading frame of *Fugu EGLN1* is highly homologous to human and mouse *EGLN1* (figure 9.4). The only feature of *Fugu EGLN1* that stands out from other *EGLN1* orthologues is the three tandem amino acid repeats of the motif "GDTPGA", located in the divergent sequence between the MYND-type finger and EGLN domain (figure 9.4). Although this motif was found in a wide range of proteins (fuzzpro, section 2.11.2), it did not occur as a tandem repeat in any of these proteins. (data not shown).

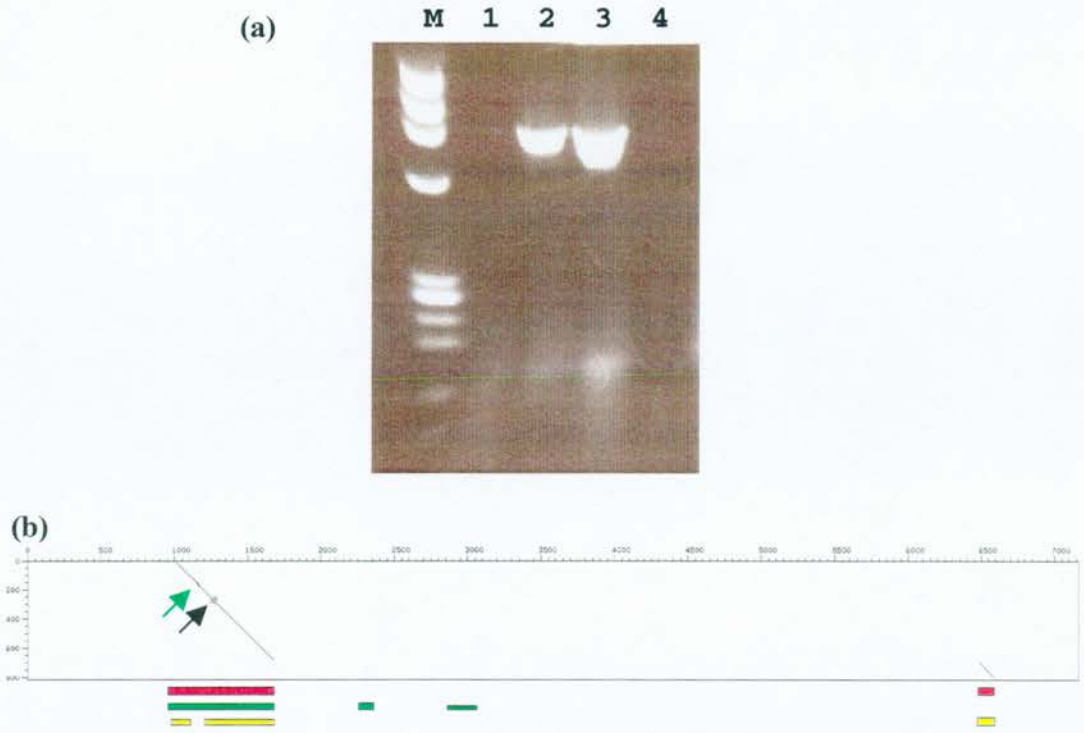


Figure 9.3; The expression and genomic structure of *Fugu EGLN1*. **(a)** RT-PCR results demonstrating the expression of EGLN1 in *Fugu* heart and ovary tissue. Lane M is the molecular weight marker, 1 is RT-PCR –ve control, 2 is *Fugu* heart cDNA template and lane 4 is *Fugu* genomic DNA template. **(b)** Dotter alignment of the partial *Fugu EGLN1* transcript (vertical) against *Fugu* genomic DNA (horizontal). The red bars below the box represents the determined structure of exons 1 and 2. The green bars indicate genscan (section 2.11.2) exon predictions. The yellow bars represent TBLASTN (section 2.11.2) HSPs when searching the *Fugu* genomic sequence with the human *EGLN1* amino acid sequence. The depth of the bars indicates the relative score of the prediction or alignment. The green arrow indicates a hexanucleotide repeat in coding sequence, that differs by one repeat between the cDNA sequence and genomic sequence. The black arrow indicates three tandem copies of an 18 nucleotide repeat within coding sequence, that encodes the GDTPGA amino acid repeat motif.

9.5 Evolutionary relationship of the EGLN genes

In global alignment, human and mouse EGLN orthologue pairs share 84%, 90% and 97% identity respectively for EGLN1, EGLN2 and EGLN3 at the amino acid level. *EGLN1* (*C1orf12*) has previously been reported to be the human orthologue of rat *Sm-20* (Dupuy *et al.,* 2000). However, it is clear from multiple sequence alignment (figure 9.5) that rat SM-20 shows substantially greater similarity with EGLN3 in aligned regions (>99% identity with mouse and 97% identity with human). The conclusion that *EGLN3* is the true mouse and human orthologue of rat *SM-20* is further supported by the observation of extended regions of homology in non-coding sequence between mouse/human *EGLN3* and *Sm-20*. No such similarity is observed between *Sm-20* and *EGLN1* or *EGLN2* (data not shown).

Human *EGLN1* (but not *EGLN2* or *3*) and *C. elegans* Egl-9 proteins both contain an N-terminal MYND type zinc finger motif (PFAM: PF01753) that shows 40% identity over 52 amino acids and conserves all of the residues critical for zinc binding (figure 9.4). This motif was conserved in *Fugu* EGLN1 and also appeared to be present in mouse EGLN1, although this sequence was incomplete at the 5' end (figure 9.4). The absence of this motif in SM-20 further supports the conclusion that *SM-20* is orthologous to *EGLN3* rather than *EGLN1*. The rat EST sequence AI510999 and overlapping cDNA fragments are strong candidates for the rat orthologue of *EGLN1* (figure 9.5).

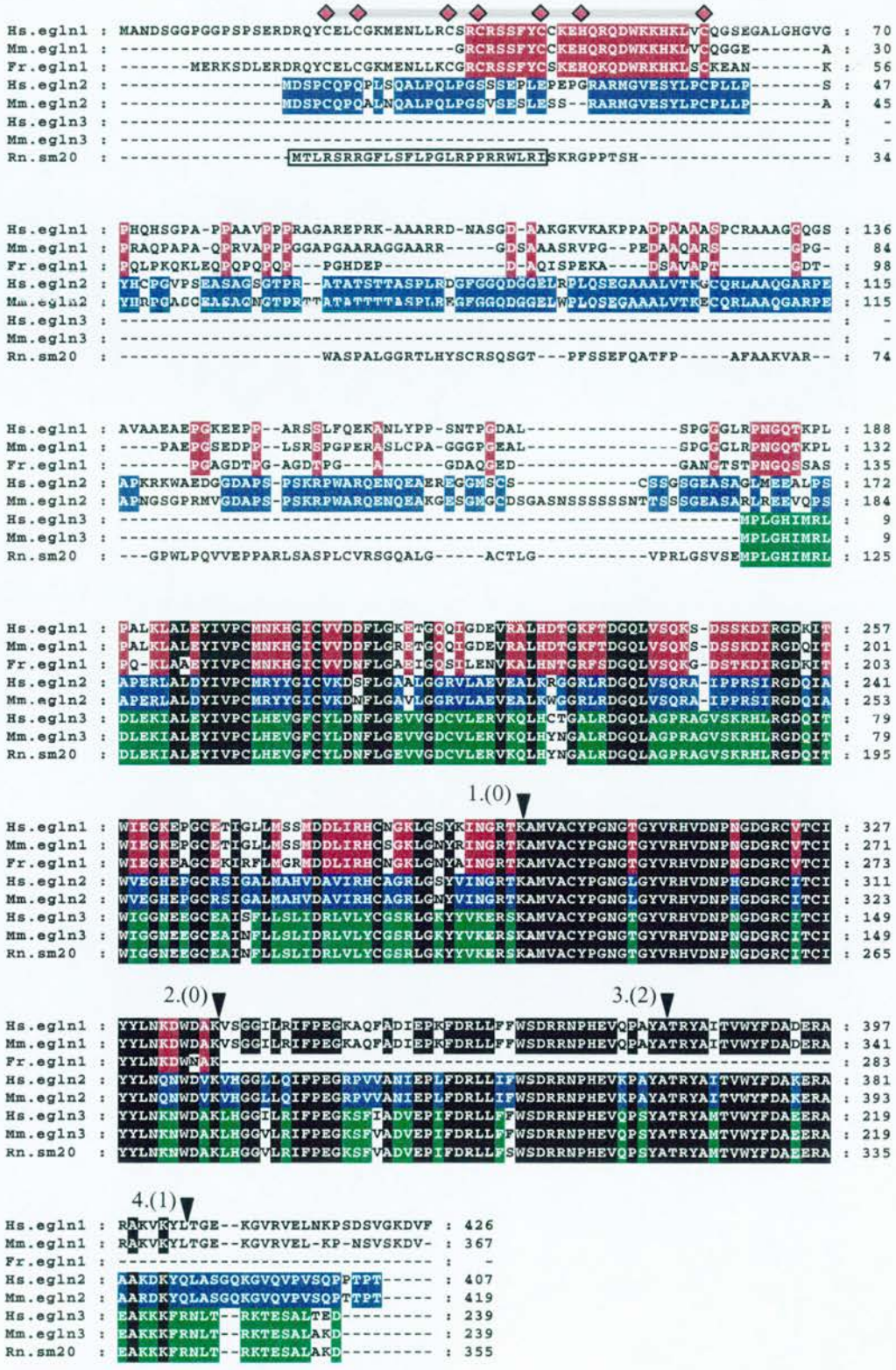


Figure 9.4; Multiple sequence alignment of vertebrate EGLN protein, amino acid sequences. For shading purposes, orthologous proteins are grouped, residues conserved between EGLN1 orthologues have a brown background, those conserved between EGLN2 proteins

have a blue background and EGLN3 protein conserved residues have a green background. Residues conserved between all EGLN family members have a black background. Only partial amino acid sequences are available for *Fugu* and moue EGLN1, shading has been configured so that the absence of sequence does not detrimentally effect the patterns of amino acid conservation. The horizontal grey bar indicates the extent of the MYND-type zinc finger in EGLN1 proteins and embedded diamonds indicate the conserved cysteine and histidine residues critical for zinc binding. The mitochondrial targeting sequence of SM-20 is boxed. Triangles above the alignment show the position of splice sites relative to the encoded amino acid sequence for all of the human *EGLN* genes, the sites of intron 1 and 2 are also conserved in *Fugu* EGLN1. Each intron is numbered, numbers in parenthesises indicate the phase of the intron relative to coding sequence.

```
Mm.egln1 GGGAGACCGGGCAGCAGATCGGCGATGAGGTGCGCGCCCTGCACGACACCGGCAAGTTCA
BF407173 ----------------------GCGGCCGCTGCCGCAGCTCCTTCTACTGCGGCAAGTTCA
                              ***        *** *      * **  ***********

Mm.egln1 CGGACGGGCAGCTGGTCAGCCAGAAGAGTGACTCTTCCAAGGACATCCGGGGGGACCAGA
BF407173 CGGACGGGCAACTGGTCAGCCAGAAGAGTGACTCTTCCAAGGACATCCGGGGGGACAAGA
         ********** ********************************************** ***

Mm.egln1 TCACCTGGATCGAGGGCAAAGAGCCCGGCTGCGAAACCATCGGCCTGCTCATGAGCAGCA
BF407173 TCACCTGGATCGAGGGCAAAGAGCCCGGCTGCGAGACCATCGGCCTGCTCATGAGCAGCA
         ********************************** ************************

Mm.egln1 TGGACGACCTGATCCGCCACTGCAGCGGGAAGCTGGGCAACTACAGGATAAACGGCCGAA
BF407173 TGGACGACCTGATCCGCCACTGCAGCGGGAAGCTGGGCAACTACAGGATAAACGGCCGAA
         ************************************************************

Mm.egln1 CGAAAGCCATGGTTGCTTGTTACCCAGGCAACGGAACAGGCTATGTCCGTCACGTTGATA
BF407173 CGAAAGCCATGGTCGCCTGTTACCCAGGCAACGGGACAGGCTACGTCCGTCACGTCGATA
         ************* ** ***************** ******** *********** ****

Mm.egln1 ACCCAAATGGAGATGGAAGATGCGTGACATGTATATATTATCTAAATAAAGACTGGGACG
BF407173 ACCCAAATGGAGACGGAAGATGTGTGACATGTATATATTATCTAAATAAAGACTGGGACG
         ************* ******** ************************************

Mm.egln1 CCAAGGTAAGTGGAGGTATTCTTCGAATTTTTCCAGAAGGCAAAGCCCAGTTTGCTGACA
BF407173 CCAAGGTGAGCGGAGGTATTCTTCGAATTTTCCCAGAAGGCAAAGCCCAGTTTGCTGACA
         ******* ** ******************** ****************************

Mm.egln1 TTGAACCCAAATTTGATAGACTGCTGTTTTTCTGGTCTGACCGGCGTAACCCTCATGAAG
BF407173 TTGAACCCAAGTTTGATAGATTGCTGTTTTTCTGGTCTGACCGGCGTAACCCTCATGAAG
         ********** ********* ****************************************

Mm.egln1 TACAGCCAGCATACGCCACAAGGTACGCAATAACTGTTTGGTATTTTGATGCAGATGAGC
BF407173 TACAGCCTGCATACGCCACAAGGTACGCAATAACCGTTTGGTATTTCGATGCACACGAGC
         ******* ************************** *********** ****** * ****

Mm.egln1 GAGCGAGAGCTAAAGTAAAATATCTAACAGGTGAGAAAGGTGTGAGGGTTGAACTCAAGC
BF407173 GAGCAAGAGCTAAAGTAAAATATCTCAC-------------------------------
         **** ******************** **
```

Figure 9.5; A candidate rat orthologue of *EGLN1*. Local sequence alignment of mouse EGLN1 (bases 481 to 1080 of embl:AJ310456) and rat EST sequence EMBL: BF407173.

9.5.1 *EGLN1* is the ancestral member of the gene family

The presence of a MYND type zinc finger in Egl-9 and its conservation in one of the vertebrate homologues suggested that the zinc finger - EGLN domain combination

represents the ancestral form of the protein, which subsequently duplicated and diverged in the lineage leading to vertebrates. Although the predicted *Drosophila* protein CG1114 a homologue of the EGLN genes (section 9.5.3), does not contain a zinc finger, the genomic sequence upstream (AE003603) has the potential to code for a MYND type zinc finger, and was predicted as a separate, single exon gene (CG14665). Based on the cross species evidence, it is likely that CG1114 and CG14665 actually represent a single gene, the *Drosophila* orthologue of *Egl-9*. This conclusion was supported by the finding that opposite end sequences of the cDNA clone LD24638 (AA820762 and AI455510) link the two predicted genes into a single transcript, when *Drosophila* ESTs were assembled in the manner described in section 2.11.4.

In addition to the characterised EGLN genes from human, mouse, rat, *C. elegans* and *Drosophila*; EST and genomic sequences from a wide variety of metazoans were identified that in translation clearly showed substantial similarity to the EGLN genes. Sequence corresponding to exon 2 of the mammalian EGLN genes was well represented amongst these sequences. After filtering for redundancy, a set of 22 homologous sequences were identified and aligned. Although this data set was composed of often poor quality sequence and its coverage of the likely EGLN gene content of genomes was certain to be incomplete, it did provide the means to investigate the minimum EGLN gene content of vertebrate genomes and the phylogeny of those genes.

The exon 2 encoded amino acids were too well conserved between homologous sequences to enable meaningful phylogenetic analysis. At the nucleotide level, using all residues for analysis produced trees that were clearly disrupted by GC3 bias (percent C+G residues in third codon positions). Using only first and second codon positions produced gene trees in good agreement with the expected relationships of human and mouse EGLN genes (section 9.5.1).

From the phylogenetic analysis (figure 9.6), it is apparent that there are *EGLN1* homologues in mammals, birds, amphibians and fish, consistent with the hypothesis

that *EGLN1* represents the ancestral member of the gene family. The bootstrap value of 39% for placing *Fugu EGLN1* and *Tetraodon* genomic survey sequence (GSS) CNS046S9 in the *EGLN1* clade is not in itself convincing. However, the conservation of synteny between *Fugu* and human *EGLN1* and the conserved domain structure (figure 9.4) strongly imply that *Fugu EGLN1* does belong in the *EGLN1* clade.

The taxonomic range of EGLN exon 2 homologues in the available sequence databases suggested that the ancestral *EGLN1* gene had duplicated and diverged by the last common ancestor of amphibians and mammals, resulting in the *EGLN2* gene. *EGLN3* was only detected in mammals and birds. Based on the analysis summarised in figure 9.6 it was not possible to determine if the complete complement of mammalian EGLN genes (*EGLN1*, *2* and *3*) reflects the complement common to all vertebrates. To address this issue, a more focussed analysis of *Tetraodon* sequences (the most completely sequenced non-mammalian, vertebrate genome to date) was carried out. Searching for *Tetraodon* genomic sequence with detectable homology to any of the human EGLN genes (TBLASTN, section 2.11.2), identified 13 GSS sequences (table 9.3). Assembly of these sequences removed redundancy and improved the quality of sequence through the generation of consensus sequences in overlap regions.

The resulting 6 contigs were then sorted on the basis of which human EGLN gene they showed greatest similarity to (table 9.3). Contigs 2 and 5, and 3 and 4 are both linked by opposite end sequences (mate pairs) of the same sub-clone. There is a perfect agreement between the ranking of EGLN gene similarity and mate pair linked contigs (table 9.3). From the clustering of *Tetraodon* sequences alone, it is apparent that fish have at least three EGLN genes with a similar pattern of homology to the mammalian *EGLN1*, *2* and *3* genes. Exons 3 and 4 are available for all three of the *Tetraodon* EGLN genes (table 9.3). Phylogenetic analysis of exon 3 and 4 coding sequence demonstrates that *Tetraodon nigroveridis* has at least three EGLN genes that are orthologues of mammalian *EGLN1*, *2* and *3* (figure 9.7).

From the high degree of exon coverage for each of the *Tetraodon* EGLN genes the relative position and phase of most exons could be determined or inferred. The structure of all three *Tetraodon* genes was consistent with the five exon, splice phase 0, 0, 2 ,1 gene structure common to human EGLN genes, with the position of introns relative to coding sequence also conserved (data not shown).

Contig number[a]	GSS sequences[b]	EGLN gene[c]	Exons[d]
1	CNS03C1C, CNS05DFD	*EGLN2*	3,4,5
2	CNS03N3R[1], CNS04IEK	*EGLN3*	2,3,4
3	CNS03QBX[2], CNS046S8[3]	*EGLN1*	3,4,5
4	CNS03QBY[2], CNS046S9[3], CNS04CEU	*EGLN1*	2
5	CNS021JC, CNS03N3S[1], CNS03Q6X	*EGLN3*	1
6	CNS04S2I	*EGLN1*	1

Table 9.3; *Tetraodon* genomic survey sequences (GSS) with homology to human EGLN genes. **(a)** Contig number is arbitrarily assigned during the assembly process. **(b)** EMBL accession numbers for each GSS. Superscript numbers after each accession number indicate sequences for which an opposite end sequence is also in the table. **(c)** Most similar human EGLN gene based on percent identity in translation. **(d)** Homologous sequence corresponding to human EGLN exons.

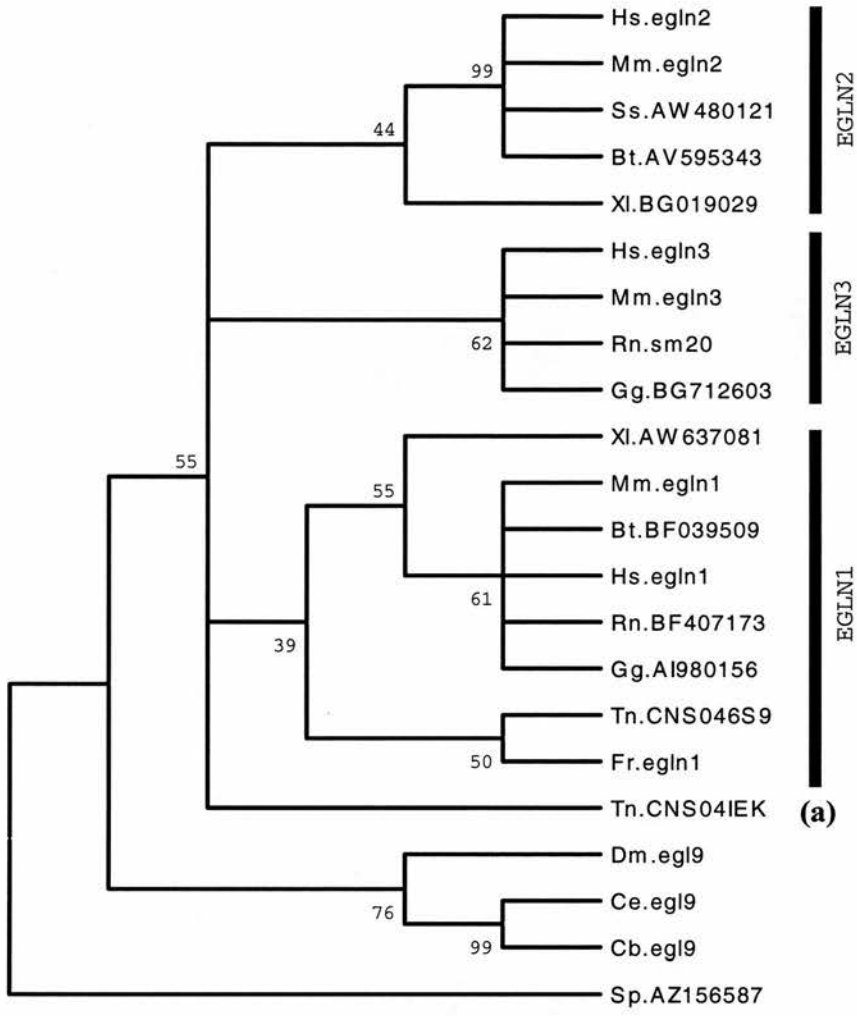


Figure 9.6; Neighbour joining, bootstrap consensus tree for aligned EGLN exon 2 nucleotide sequences. Only the first and second codon positions were considered in constructing the tree. Numbers at branch forks indicate the percentage bootstrapping support (the number of times the group consisting of sequences to the right of that fork occurred among the trees out of 1000 replicates). Sequences are named using initials of the genus and species followed by the gene name (where the gene has been previously described) or sequence accession number. Genus and species abbreviations are: Rn, *Rattus norvegicus*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Gg, *Gallus gallus*; Bt, *Bos taurus*; Xl, *Xenopus laevis*; Tn, *Tetraodon nigroveridis*; Fr *Fugu rubripes*; Ss, *Sus scrofa*; Sp, *Strongylocentratus purpuratus* (sea urchin); Ce, *Caenorhabditis elegans*; Cb, *Caenorhabditis brigsiae;* Dm, *Drosophila melanogaster*. **(a)** The sequence Tn.CNS04IEK is a *Tetraodon* genomic survey sequence that contains a frame shift, altering the EGLN reading frame. For the purposes of this analysis, that frame shift was corrected by the insertion of an ambiguous residue.

(a)

```
Hs.egln1 : RIFPEGKAQFADIEPKFDRLLFFWSDRRNPHEVQPAYATRYAITVWYFDADERARAKVKYLTG--EKGV : 410
Tn.egln1 : RIFPEGKAQFADIEPKFDRLLFFWSDRRNPHEVQPAYATRYAITVWYFDGDERARAKEKYLTR--ERGV : 112
Mm.egln1 : RIFPEGKAQFADIEPKFDRLLFFWSDRRNPHEVQPAYATRYAITVWYFDADERARAKVKYLTG--EKGV : 354
Hs.egln3 : RIFPEGKSFIADVEPIFDRLLFFWSDRRNPHEVQPSYATRYAMTVWYFDAEERARAEAKKKFRNL--TRKT : 232
Mm.egln3 : RIFPEGKSFVADVEPIFDRLLFFWSDRRNPHEVQPSYATRYAMTVWYFDAEERARAEAKKKFRNL--TRKT : 232
Tn.egln3 : RIFPEGKAGAADVAPLFDRLLFFWSDRRNPHEVQPSYSSRYAITVWYFDSVERAEAKRRFRAI--RGKT : 230
Hs.egln2 : QIFPEGRPVVANIEPLFDRLLIFWSDRRNPHEVKPAYATRYAITVWYFDAKERAAAKDKYQLASGQFGV : 396
Mm.egln2 : QIFPEGRPVVANIEPLFDRLLIFWSDRRNPHEVKPAYATRYAITVWYFDAKERAAARDKYQLASGQKGV : 408
Tn.egln2 : ---PEGKNVVASIEPLFDRLLIFWSDRRNPHEVKPAYATRYAITVWYFDAKERAEAKEKYKLATGQKGV :  66
```

(b)

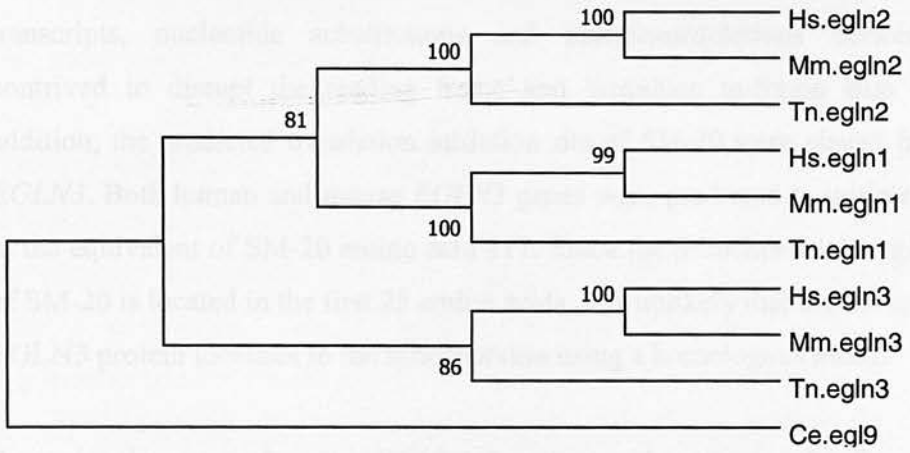


Figure 9.7; Phylogenetic analysis of vertebrate EGLN genes. Hs indicates *Homo sapiens*; Mm, *Mus musculus* and Tn, *Tetraodon nigroveridis*. **(a)** Multiple sequence alignment of EGLN gene exon 3 translations. Black background indicates 100% residue conservation, dark grey 80% identity and light grey 60% identity. Numbers to the right of the alignment show the amino acid coordinate (relative to available sequence) of the last residue in the alignment. **(b)** Neighbour joining (p-distance), bootstrap consensus tree of the alignment shown in panel a, with the equivalent *C. elegans* Egl-9 amino acid sequence added as an out group to root the tree. Numbers at branch forks indicate the percentage bootstrapping support out of 1000 replicates.

9.5.2 Mitochondrial targeting is not a general feature of the gene family

Rat SM-20 has been demonstrated experimentally to localise to the mitochondria. This localisation is dependent on the first 25 amino acids of the protein (Lipscomb *et al.,* 2001). It was also noted by Lipscomb et al. that the mitochondrial localisation was not necessary for the role of SM-20 in the induction of apoptosis. Neither mouse nor human *EGLN3* were predicted to encode amino acids corresponding to the first 116 amino acids of SM-20. While there was nucleotide similarity between the 5' coding sequence of *SM-20* aligned to the 5' UTR of human and mouse *EGLN3* transcripts, nucleotide substitutions and insertions/deletions between species contrived to disrupt the reading frame and introduce in-frame stop codons. In addition, the predicted translation initiation site of *SM-20* were absent from human *EGLN3*. Both human and mouse *EGLN3* genes were predicted to initiate translation at the equivalent of SM-20 amino acid 117. Since the mitochondrial targeting signal of SM-20 is located in the first 25 amino acids, it is unlikely that the mouse or human EGLN3 protein localises to the mitochondria using a homologous motif.

Screening human and mouse EGLN1-3 amino acid sequences for the presence of mitochondrial targeting sequences (MITDISC, Nakai and Kanehisa, 1992) failed to detect any such signals. The N-terminal mitochondrial targeting signal for SM-20 was well predicted by this method (MITDISC score of 5.29). Through C-terminal GFP fusion, Darby et al (1999) demonstrated that *C. elegans* Egl-9 localised predominantly to the nucleus with some signal retained in the cytoplasm. Combined, these results demonstrate that mitochondrial targeting is not a general feature of this gene family and may be a property peculiar to rat *SM-20*. This finding brings into question the relevance of SM-20 mitochondrial localisation in the general model of neuronal apoptosis proposed by Lipsomb *et al.,* (2001) (figure 9.1).

9.5.3 Definition and function of the EGLN domain

A hidden Markov model of the conserved EGLN C-terminal domain (subsequently the EGLN domain) was generated from amino acid alignments (figure 9.4) and used in conjunction with BLAST and Genewise (sections 2.11.2) to search for other

members of the EGLN domain containing family of proteins. In *C. elegans* genomic sequence, the only family member identified was *Egl-9*. In human genomic and EST sequences, all detected similarity to the EGLN domain was accounted for by the *EGLN* 1-3 genes, *SCAND2*, and the *EGLN3* processed pseudogene located on chromosome 12 (table 9.1). Considering the >90% sequence coverage of the human genome and depth of EST coverage for *EGLN* 1-3, it is likely that these 3 genes represent the total complement of this gene family in the human genome. A single EGLN domain containing gene has been predicted from the *Drosophila* genomic sequence, the hypothetical 325 amino acid protein CG1114 (SPTR:Q9VN98). This *Drosophila* gene, as predicted, would encode a C-terminal EGLN homology domain and a less conserved N-terminal region (see section 9.4.1).

Aravind and Koonin (2001) have provided evidence that the conserved EGLN domain comprises a sub-family of the 2OG-Fe(II) oxygenase superfamily (2-oxoglutarate and iron II dependant oxygenases). This super family of proteins is widely distributed through eukaryotes and eubacteria. The *Streptomyces ansochromogenes* protein SANF (SPTR: Q9KIT9) was also identified as a member of this protein superfamily (Aravind and Koonin, 2001). SANF was the highest scoring alignment with the EGLN domain HMMs (section 2.11.2), other than the identified EGLN family proteins (hmmsearch score of 18.0 versus 549.9 to 406.8 for EGLN family proteins). On this basis, the SANF amino acid sequence was used as an out group for phylogenetic analysis of the EGLN conserved domain (section 9.5).

Based on the determined crystal structures of other members of the 2OG-Fe(II) oxygenase super family (Zhang *et al.,* 2000; Valegard *et al.,* 1998 and Roach *et al.,* 1995), two histidine and an aspartic acid residue are critical for binding the Fe(II) group and a further two residues (R279 and F285 in *Emericella nidulans* isopenicillin N synthase structure; PDB:lips) contribute to the catalytic site of this super family of proteins (figure 9.4). With the exception of *Emericella nidulans* isopenicillin N synthase residue F285, all of the catalytic residues were conserved in the ELGN domain. The position equivalent to F285 in EGLN proteins was substituted for tryptophan, a conservative substitution that was also found in YbiX,

leprecan, prolyl hydroxylase and members of the SanF/SanC families of 2OG-Fe(II) oxygenase proteins (Aravind and Koonin, 2001). On the basis of this structural information and the conservation of catalytic residues, Aravind and Koonin have proposed the Egl-9 subfamily (EGLN domain) of 2OG-Fe(II) oxygenase folds are amino acid hydroxylases that modify protein targets.

9.5.4 Horizontal gene transfer?

Searching of the EST and genomic sequence databases identified strong homologues of the EGLN sub-family in many vertebrates, *D. melanogaster*, *C.elegans* and the ascidian *Ciona intestinalis* (AV675485 & AV680334), all of which are metazoan eukaryotes.

It is striking that given the apparent metazoan origin of this gene family, genes with the potential to encode proteins clearly of the EGLN sub-family were identified in the pathogenic bacteria *Pseudomonas aueruginosa* and *Vibrio cholerae*, (figure 9.8; Aravind and Koonin 2001; Taylor 2001b). Close homologues of the *P. aueruginosa* EGLN related gene (Q9I6I1) are also present in the unfinished bacterial genome sequences of *Pseudomonas syringae*, *Pseudomonas putida* and *Shewanella putrefaciens* (http://www.tigr.org/) (figures 9.8 and 9.9). *Vibrio*, *Psuedomonas* and *Shewanella* are all members of the gamma subdivision of proteobacteria. However, the EGLN sub-family appears absent (based on HMM searching) from the genomes of the other fully and partially sequenced proteobacteria, archae, yeast, plants and viruses. It is particularly interesting that other proteobacteria of the gamma subdivision (including *E. coli*) appear to lack this domain.

This mosaic pattern of phylogenetic distribution is most readily explained by horizontal gene transfer from metazoans to bacteria (Aravind and Koonin, 2001). However the other possibility that must be considered is vertical inheritance and selective loss in many lineages, with specific retention in metazoans and some proteobacteria. Phylogenetic analysis using *Streptomyces ansochromogenes SANF* amino acid sequence as an out group (figure 9.9) shows that all bacterial EGLN domains cluster together and all metazoan EGLN domains cluster together. The convincing separation (100% bootstrapping support) of these two clades supports a

vertical inheritance model rather than horizontal transfer between metazoans and eubacteria. *C. elegans* Egl-9 has previously been directly implicated in host pathogen interactions with *P. aeruginosa* (Darby *et al.,* 1999 & Johnson and Liu, 2000) . The presence of an unexpectedly close homologue in *P. aeruginosa* and other pathogenic bacteria suggests the EGLN domain may be of significance to pathogenicicty (Taylor 2001b), although a mechanism for this remains unclear.

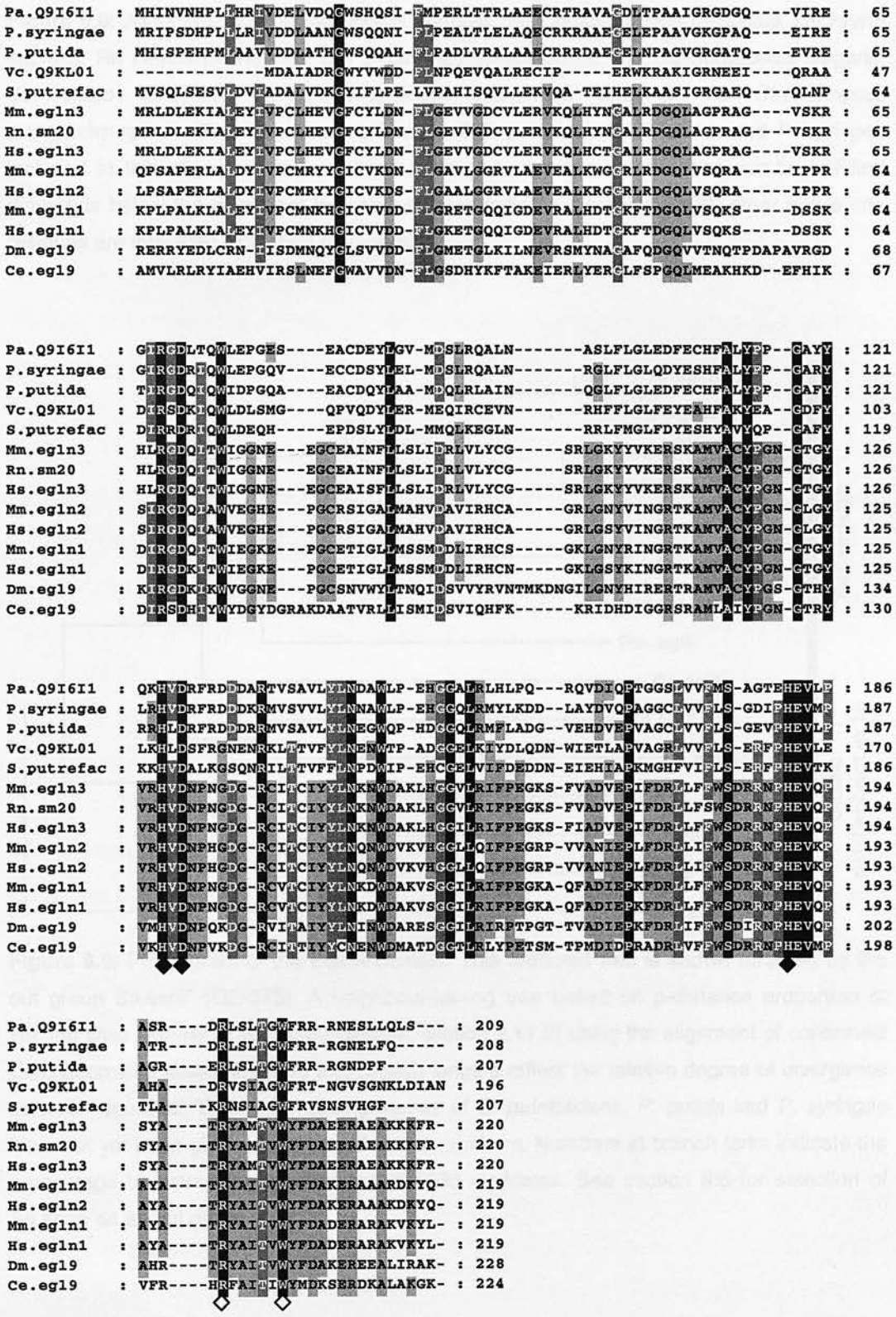


Figure 9.8; Alignment of all known EGLN domains. Full legend on next page.

Figure 9.8; Alignment of all known EGLN domains. Mm indicates *Mus musculus*, Hs *Homo sapiens*, Rn *Rattus norvegicus*, Dm *Drosophila melanogaster*, Ce *Caenorhabditis elegans*, Vc *Vibrio cholerae*, Pa *Pseudomonas auerginosa* and Sa is *Streptomyces ansochromogenes*. The open reading frames of *S. putrefaciens*, *P. putida* and *P. syringae* included in this alignment have not yet been given names or reference numbers. Filled diamonds below the alignment indicate residues critical for binding Fe(II), other active site residues are indicated with open diamonds.

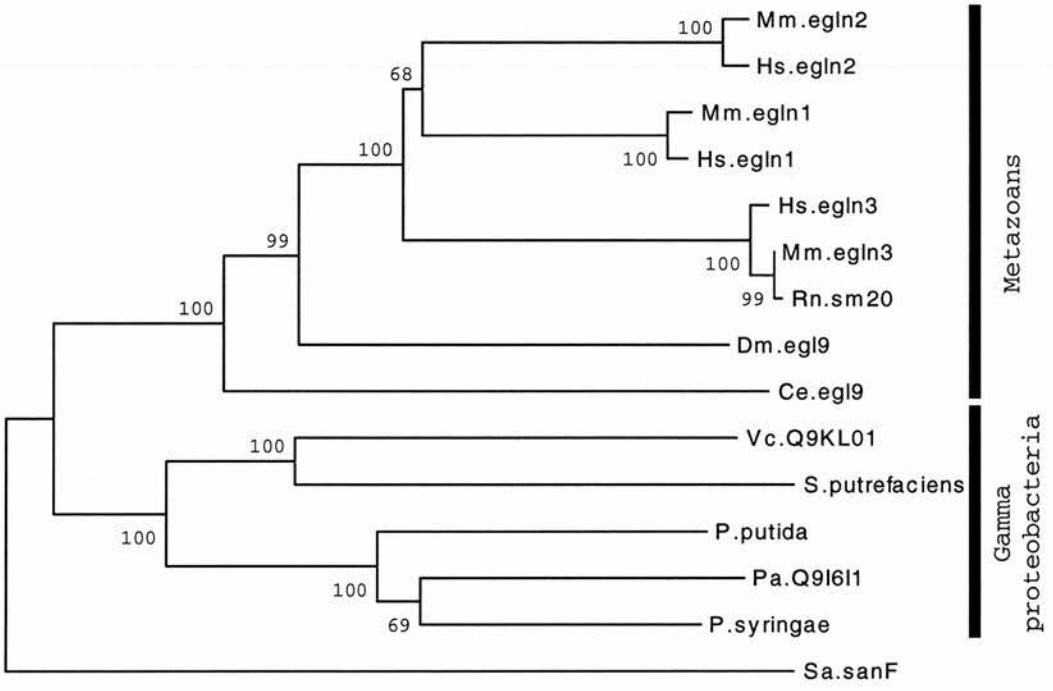


Figure 9.9; Phenogram of the EGLN domain. This unrooted tree is shown rerouted by the out group Sa.sanF (Q9KIT9). A neighbour-joining tree based on p-distance proportion of varying sites (Kumar et al., 2001) (Mega2, section 2.11.2) using the alignment of conserved EGLN domains shown in figure 9.8. Branch lengths reflect the relative degree of divergence between domains. The open reading frames of *S. putrefaciens*, *P. putida* and *P. syringae* have not yet been given names or reference numbers. Numbers at branch forks indicate the percentage bootstrapping support out of 1000 replicates. See section 8.5 for selection of Sa.sanF as an out group.

9.6 Discussion

The work presented in this chapter has described human *EGLN1, EGLN2* and *EGLN3*, and mouse *EGLN1*, *EGLN2* and *EGLN3* genes for the first time, as well as characterising the wider gene family and its evolution. Human, mouse and *Tetraodon* genomes encode 3 homologues of the *C. elegans Egl-9* gene. Each of these homologues have a conserved gene structure suggesting duplication in the lineage leading to vertebrates, or early in the vertebrate lineage prior to the last common ancestor of bony fish and mammals.

Based on patterns of amino acid conservation and domain organisation, *EGLN1* represents the ancestral form of this gene family in metazoans. Contrary to previous reports (Dupuy *et al.,* 2000; Lipscomb *et al.,* 1999), rat *Sm-20* was demonstrated to be the orthologue of human and mouse *EGLN3* rather than *EGLN1* (*C1orf12*). The lack of potential mitochrondrial targeting signals in *EGLN3* and the observation that mitochondrial targeting is not necessary for the role of *SM-20* in caspase mediated apoptosis, suggest that mitochondrial targeting is not a general feature of the *EGLN* gene family and is likely to be a peculiarity of rat *SM-20*.

The presence of *EGLN*-like genes in *Gamma Proteobacteria*, and their absence in other bacteria including some *Gamma Proteobacteria*, suggests a highly lineage specific retention of this gene family, although horizontal transfer cannot be ruled out. The known role of *Egl-9* in the susceptibility of *C. elegans* to a *P. aeruginosa* toxin combined with the presence of an EGLN gene in this bacteria is an intriguing coincidence, further study of which may provide new insight into host pathogen interactions.

In the human genome, *EGLN1* maps to chromosome 1q42, *EGLN2* to 19q and *EGLN3* to 14q. *SCAND2* has previously been reported to map to chromosome 15 and there is an *EGLN3* processed pseudogene on chromosome 12p. The *EGLN2* locus on chromosome 19q.13.13 is approximately 5 Mb from the APOE gene at 19q13.2 according to ENSEMBL 1.1.1 (section 2.11.1). There has been one, unreplicated

report of schizophrenia association with the APOE, E4 allele (Harrington *et al.,* 1995). There is no strong evidence for a major mental illness susceptibility locus around the *EGLN3* gene on chromosome 14. The general lack of a correlation between ELGN genes and linkage or association hotspots in the genome suggests that EGLN genes as a whole are not good positional candidates for susceptibility to major mental illness.

The EGLN genes are likely to function enzymatically as 2-oxoglutarate and Fe (II)-dependant oxygenases, probably targeting polypeptides for hydroxylation (Aravind and Koonin, 2001). EGLN3 orthologues are directly implicated in the regulation of cell growth and apoptosis (section 9.2), with a specific role in nerve growth factor dependant survival of neurons (Lispcomb *et al.,* 1999 and Lispcomb *et al.,* 2001). The growth factor dependant survival of neurons is an important mechanism in neural patterning, a process that has been reported to be disrupted in some schizophrenics (Woods, 2000 for review). It is not currently known if this is an anatomical feature of t(1;11) carriers although it could provide an explanation for the increased latency in P300 ERP seen in t(1;11) carriers (section 1.6). An implicated role in neural patterning is sufficient to consider *EGLN1* a plausible functional candidate for involvement in the genetic aetiology of major mental illness.

A toxin acting through Egl-9 is able to induce dramatic and lethal neuromuscular paralysis of *C. elegans* (Derby *et al.,* 1999) at speeds too rapid to be explained by regulation of growth or apoptosis. This suggests that EGLN genes, particularly *EGLN1* which is considered an orthologue of *Egl-9* have an additional role in neural signalling. This provides a second reason to consider *EGLN1* a functional as well as a positional candidate in the aetiology of major mental illness. Future work could build on these findings and test *EGLN1* specifically in association studies for its potential contribution to the susceptibility of major mental illness.

Chapter 10

Discussion and future directions

10.1 Summary

The work presented in this thesis has investigated the genomic sequence and genes around a chromosome translocation breakpoint that is strongly implicated as a susceptibility locus for major mental illness. At the time of initiating this work, the partial sequence for the predicted protein coding gene *DISC1* and a second, anti-parallel transcript *DISC2* had been identified. The structure and function of the *DISC1* gene product was unknown and sequence for a homologue of *DISC1* had not been described. The nature of the *DISC2* transcript was undetermined. The wider genomic context of the translocation was unknown.

Results presented in chapters 3 to 9 have culminated in several lines of work that substantially increase understanding of the breakpoint locus and provided resources for the further investigation of the locus. Contiguous human genomic sequence was assembled over the region and the orthologous region from *Fugu rubripes* was identified and sequenced. The human transcription map was refined and a transcription map of the *Fugu* locus constructed. Comparative genomic analysis demonstrated conservation of synteny between the *Fugu* and human loci and lead to the identification of *EGLN1*, a novel protein coding gene. Homologues of the *DISC1* gene were identified from a range of vertebrates, analysis of these sequences resulting in a generalised model of the *DISC1* protein structure and defined regions of protein – protein interaction. *TRAX* and *EGLN1* were demonstrated to be the closest protein coding genes to the translocation breakpoint that were not directly disrupted by it. Both *TRAX* and *EGLN1* were investigated and new protein families were described as a result of those investigations. The apparently non-protein coding transcripts *DISC2*, *Backtrax* and *Foretrax* were not evolutionarily conserved neither was the intergenic splicing of *TRAX* to *DISC1*.

10.2 Comparative genomic analysis

Initial cloning of the *DISC1* genomic region from *Fugu* was unsuccessful due to the poor conservation of this gene between species. Problems of poor conservation were confounded by initial speculation that the N-terminal region represented largely by exon 2 would be more conserved than the coiled coil containing C-terminal region of the protein. The finding that the *TRAX* gene was likely to be directly upstream of *DISC1* (Millar *et al.,* 2000a) provided a new opportunity to identify the orthologous region from *Fugu*. This strategy relied on the conservation of synteny between humans and *Fugu*. The *TRAX* gene was rapidly identified by cross-species hybridisation and a homologue of *DISC1* identified by sequence sampling the *TRAX* containing clones. Forty five kiliobases of contiguous genomic sequence was produced from the clone contig. RT-PCR and cDNA library screening proved to be efficient means of constructing a *Fugu* transcription map and validating gene structure prediction. Clones representing the orthologous region of the chicken genome were isolated in the manner used to identify the *Fugu* region, providing a resource for future comparative genomic investigation of the locus.

The impact of the human genome sequencing project and other genome and EST sequencing projects on this work has been enormous. Although for much of the human genome, sequence remains in a draft status, the relative order and orientation of sequence fragments can often be inferred from sequence and clone overlaps. Attempts at whole genome sequence assembly have met with mixed success (Semple *et al.,* in press) and are likely to remain extremely error prone until the majority of the genome is represented by finished BAC clone sequence (Eichler, 2001). However, it was demonstrated that for localised regions of the genome, a high accuracy assembly could be constructed by combining the complementary draft mapping and sequence information from both the International Human Genome Sequencing Consortium (IHGSC) and Celera Genomics sequence (section 4.3.2). The generation of a 0.8 Mb sequence contig (1 gap) over the breakpoint region provided a key resource for the annotation and comparative analysis of this locus.

The assembled human genomic sequence and contiguous *Fugu* sequence was subjected to preliminary annotation (chapter 5) by *ab initio* gene prediction methods and sequence similarity searching against known sequences. This preliminary annotation was sufficient in every case examined, to indicate the presence of a protein coding gene. However, the finer details of genomic structure, extent of UTR and alternative splicing could not be resolved in the majority of cases by preliminary annotation alone. Multiple EST clusters within the introns of *DISC1* were identified and subsequent experimental work demonstrated that they represented splice variants of the *DISC1* gene (chapter 6). The preliminary annotation also indicated the existence of the *Backtrax* and *Foretrax* transcripts. These transcripts were not obviously protein coding but were represented by multiple cDNA clones from multiple libraries indicating that like *DISC2* they do represent genuine transcripts.

Comparison between the human and *Fugu* genomic sequences and transcription maps demonstrated that the *EGLN1*, *TRAX* and *DISC1* genes were conserved in order and orientation between species (section 4.2). Human *DISC1* exon 11 was found to be represented by two exons in *Fugu* (11 and 11a), suggesting the loss of an intron in the lineage to mammals. Interestingly, humans have retained the ability to alternatively splice the distal region of exon 11 sequence that represents exon 11a in *Fugu*. The retention of this ability to alternatively splice exon 11a even after exon fusion suggests that the alternative splicing is of biological significance. This relatively subtle change in genomic structure has dramatic consequences on the splicing strategies used to produce equivalent protein isoforms (section 6.3.2). In addition to the alternative inclusion or exclusion of the exon 11 distal sequence, several C-terminally truncated *DISC1* splice variants have been identified. In humans, alternative terminal exons and 3' UTRs are utilised to truncate the *DISC1* open reading frame shortly after exon 9 coding sequence and a further alternative terminal exon results in termination of the reading frame shortly after exon 3 (section 6.5.2). A *Fugu* equivalent of the post exon 3 truncation (E3 isoform) was not identified. However, two *Fugu* splice variants were identified that splice from exon 9 to exon 11a or exon 12. Both of these splicing events would shift the

downstream reading frame and cause a truncation of the reading frame shortly after exon 9 coding sequence. The conserved extent of open reading frame, but altered splicing strategies again suggests that these splice variants are biologically significant.

10.3 The *DISC1* gene product

The hypothetical protein encoded by *DISC1* was novel, the only significant sequence similarity with other proteins was with myosins and other proteins with extended stretches of coiled coil (section 3.2.1). This sequence similarity was suggestive of a coiled coil structural feature but not necessarily of homology to myosins or other coiled coil containing proteins (section 3.2.1). The absence of clear *DISC1* homologous sequences prohibited meaningful analysis of the DISC1 amino acid sequence as it would be unclear if detected features reflected the chance occurrence of a recognised sequence motif or a biologically relevant feature. To address this issue, the complete open reading frame for the longest splice variant of *DISC1* was demonstrated for *Fugu*, predicted for mouse and the majority of the open reading frame for a zebrafish homologue of *DISC1* was also determined. These additional sequences allowed the properties of *DISC1* products to be evaluated.

The most prominent feature of DISC1 homologues was the conserved modular nature of predicted coiled coil regions in the C-terminal half of the protein (section 6.7.2). This is in marked contrast to the single continuous block of coiled coil found in myosins and other force transducing molecules. The individual blocks of coiled coil could interact with discrete protein targets suggesting that DISC1 may be involved in mediating multiple protein – protein interactions. Accordingly, the various C-terminally truncated forms of DISC1 would have alternate combinations of these coiled coil blocks, particularly the well conserved C-terminal most coiled coil region. Consistent with this prediction, preliminary findings from yeast-2-hybrid investigations have identified two proteins that are predicted to have coiled coil forming regions that interact strongly with full length DISC1, but not a truncated version that only contains coding sequence represented by exons 1 to 9 (K. Millar, personal communication).

The N-terminal region of DISC1 was found to be poorly conserved between all species, although the general compositional bias of the sequence was maintained (section 6.7.1). Two small N-terminal motifs were found to be highly conserved between DISC1 homologues, specifically an arginine rich motif close to the N-terminus of the protein and a serine and phenylalanine rich motif more centrally located in the N-terminal region. The arginine rich motif conforms to a known consensus nuclear localisation signal (section 6.7.1). The serine and phenylalanine rich motif does not conform to any known conserved motif. The specific conservation of a nuclear localisation signal within the background of poor conservation suggests that the arginine rich motif represents a functionally conserved nuclear localisation signal.

Although a generalised structure and probable nuclear localisation could be predicted, it was not possible to ascribe a specific function to the DISC1 protein. However, the finding of discrete blocks of coiled coil forming potential in the C-terminal region and a consistent compositional and predicted structural distinction between the N-terminal and C-terminal regions suggest functionally discrete "modules" of the protein. Such modules could be tested in isolation for protein – protein interactions and other subsequent functional studies.

The modular nature of DISC1 also has bearing on the possible consequences of the translocation. Northern blot analysis indicates the 7.5 kb transcript representing the longest *DISC1* isoforms (L1 and/or L2) is the most abundant *DISC1* transcript in the cell types tested, including the brain (Millar *et al.,* 2000). Within translocation carriers, the derived chromosome 1 could theoretically transcribe the shortest known *DISC1* splice variant (the E3 variant) and/or could also produce transcripts containing coding sequence up to exon 8. Such transcripts would not encode the C-terminal most coiled coil forming regions, potentially sequestering proteins that normally interact with DISC1 and undermining the action of the full length L1 and L2 DISC1 products derived from the non-translocation chromosome 1. Through such a mechanism a truncated form of DISC1 could exert a dominant negative

influence. This could explain the dominant mode of inheritance exhibited by the susceptibility to major mental illness segregating through the t(1;11) family.

10.4 Non-coding mRNA like transcripts

Polyadenylated transcripts that are either spliced or unspliced are generally expected to be protein coding. Those mRNA like transcripts that do not appear to be protein coding are considered rare and novel (Erdmann *et al.,* 1999). Some clearly have function such as the *XIST* transcript involved in mammalian sex chromosome dosage compensation (Avner and Heard, 2001 for review). However, the majority of these transcripts are without known function (Erdmann *et al.,* 2001). *DISC2* appears to represent this group of non-coding mRNA like transcripts, principally because of the lack of a substantial open reading frame. Other factors such as the length of the transcript and the distribution of interspersed repetitive elements suggest that it is unlikely to encode a protein (section 7.2).

Similarly, the *Backtrax* and *Foretrax* transcripts were mRNA-like but did not encode substantial open reading frames. Comparative sequence analysis indicated that none of the short open reading frames present in the transcripts were conserved between humans and *Fugu.* Neither was there evidence of conservation in the genomic sequence or EST data sets from other mammals. The close proximity of the *Backtrax* transcription start site to the *TRAX* transcription start site suggests that these transcripts share a common promoter. A possible explanation for the existence of the *Backtrax* transcript is the "leaking" of the *TRAX* promoter causing the RNA polymerase to be loaded on the wrong DNA strand and transcribed away from the *TRAX* gene. Cryptic sites in unique sequence and interspersed repetitive elements would then be utilised as splicing and polyadenylation signals by the aberrant transcript. The existence of bi-directional promoters is well established (Masumoto *et al.,* 2001). Bi-directional "leaking" is often associated with the insertion of a promoter carrying transgene (Scacheri *et al.,* 2001; Phan *et al.,* 2001) indicating that many promoters may be intrinsically bi-directional and additional insulator elements are required to confer directionality.

A good candidate for the promoter activity of the *DISC2* transcript was an endogenous retrovirus integrated in intron 9 of *DISC1* (chapter 7). The pro-retroviral sequence possessed two apparently intact TATA box containing promoters oriented in an appropriate manner to direct the transcription of *DISC2*. While transcripts could be detected by RT-PCR between known components of the *DISC2* transcript and the retroviral element, it was not possible to unequivocally distinguish the *DISC2* transcript from *DISC1* hnRNA. Consequently, the retroviral element remains an excellent candidate promoter for *DISC2* and further work is required to define the 5' end and transcriptional origin of this transcript. Strand specific nucleotide hybridisation onto Northern blots would distinguish *DISC2* transcripts from *DISC1* and allow the 5' end of the transcript to be mapped.

It is a dogma of molecular biology that DNA is transcribed to RNA, and RNA is translated to protein. Therefore the function of a gene is manifest as a polypeptide. This is clearly not always the case. For example, transfer RNAs, small nuclear RNAs (involved in splicing) and ribosomal RNAs all have biological functions as RNA end-products. However, the function of non-protein coding RNAs that do not fall into one of these categories cannot readily be investigated with current paradigms of experimental gene function analysis. Non-coding RNAs have frequently been associated with genomic imprinting (section 7.3). While this association likely reflects a substantial acquisition bias, it did raise the possibility that the t(1;11) phenotype may represent a disruption of imprinting at the breakpoint locus. *DISC2* was found to be biallelically expressed in foetal heart tissue (section 7.3), the principal known site of *DISC2* expression. This result indicates that *DISC2* is not subject to imprinting. It remains a possibility that *DISC2* is imprinted in a sub-set of tissues or developmental stages although there is no evidence that this is the case. A further possible role for *DISC2* is the transcriptional or post-transcriptional regulation of *DISC1* expression through antisense interference. However, by Northern blot analysis (Millar *et al.,* 2000a), there is no recognisable correlation between steady state levels of *DISC1* and *DISC2* transcripts. The comparison of DISC1 protein levels with *DISC2* transcript levels between cell types awaits development of specific antibodies to the DISC1 protein.

The lack of cross-species conservation and apparent low levels of transcription for the *DISC2*, *Foretrax* and *Backtrax* transcripts (sections 7.1 and 5.2.2), suggest that these apparently non-coding transcripts may represent rare aberrant transcripts without significant biological function. Further insight into the nature and function of these transcripts may be gained from greater depth of sequences for comparative analysis including a range of species more closely related to humans such as ruminants, monkeys and primates. It would also be interesting to investigate the sub-cellular localisation of the transcripts by fluorescent *in situ* hybridisation. For example XIST localises to the inactive X chromosome (Avne and Heard, 2001) and the BC1 non-coding RNA localises to neural dendrites (Muramatsu *et al.,* 1998). In both cases, localisation indicates function.

10.5 Evaluation of functional candidacy

As well as a karyotypical marker, the reciprocal chromosomal translocation that co-segregates with mental illness in the t(1;11) family is a good candidate for the disease causing aberration. However, the molecular basis of disease susceptibility in this family may not be through the direct disruption of a transcription unit by the translocation, but could be due to exerting long range position effects on transcriptional regulation (section 1.6.2). A further possibility is that the t(1;11) translocation is in linkage disequilibrium with the causative molecular aberration (section 1.6.3). For these reasons it was necessary to consider the functional candidacy of genes near the breakpoint for involvement in the observed phenotype of translocation carriers. Good functional candidates could then be prioritised for association studies and further biological investigation.

The major problem in evaluating functional candidacy of genes for a role in the aetiology of mental illness is the lack of understanding of the disease process. For example, a gene whose protein product is involved in heme biosynthesis would not generally be considered a strong functional candidate for susceptibility to psychosis. However, the disease variegate porphyria (OMIM: 176200) which is associated with schizophrenia like episodic psychotic episodes is caused by mutations in the *PPOX* gene (Deybach *et al.,* 1996). *PPOX* encodes an enzyme expressed in liver and

erythroid progenitor cells that is involved in the heme biosynthetic pathway. It is the accumulation of heme progenitors that induces psychosis (Brenner and Bloomer, 1980). This is not thought to represent a general mechanism leading to psychosis but does illustrate the indirect means by which a gene may contribute to mental illness susceptibility.

While the only limits on defining a gene as a functional candidate can be considered to be the limits of a researchers' imagination, there are clues to the molecular nature of schizophrenia susceptibility. In particular, brain structural abnormalities (section 1.2.3) combined with the biological markers associated with schizophrenia (section 1.2.4) suggest that in many cases schizophrenia represents a neurodevelopmental disorder. Specifically for the t(1;11) family, the increased latency and decreased amplitude of P300 ERPs (section 1.2.4) suggests that there is an underlying neurological abnormality in translocation carriers, independent of psychiatric diagnosis. Therefore, stronger functional candidates are those with a demonstrable neurological function, particularly neurodevelopmental and neural signalling roles.

The closest protein coding genes to the chromosome 1 translocation breakpoint were *DISC1*, *TRAX* and *EGLN1*. Each of these genes was considered in detail and new insight into the function and evolution of each of the genes was gained. The functional candidacy of *DISC1* was discussed previously (section 10.2). The *TRAX* and *EGLN1* genes were both argued to be rational functional candidates contributing to the t(1;11) phenotype. *TRAX* regulates dendritic localisation and translational masking of specific mRNAs via interaction with the Translin protein. *EGLN1* has a previously demonstrated role in regulating neuronal survival (section 9.2) and an implicated role in neuronal signalling (section 9.6).

In the wider genomic context, the genes *GNPAT, EXO84, KIAA1389, KIAA1383, DJ876B10.3, Q9BSD7* and *PCNXL2* were identified. Of all of these genes, only *PCNXL2* was considered a "good" functional candidate. *PCNXL2* is a homologue of the *Drosophila Pecanex* gene that is crucial for neurodevelopment and neural patterning (section 4.2.9 for review). This annotation alone would warrant the

consideration of *PCNXL2* as a functional candidate. Further interest is raised by the observation of striking phenotypic similarities between *Drosophila Pecanex* and *Notch* mutants (LaBonne *et al.,* 1989; section 4.6 for review). The *Notch* signalling pathway has been repeatedly implicated in the molecular aetiology of human behavioural disorders including mental illness and Alzheimer disease (OMIM: 190198; OMIM: 104300). Most notably, the *Dishevelled* gene (*Dvl1*) the protein product of which it is an antagonist of NOTCH signalling has been proposed as one of the most convincing mouse genetic models of mental illness to date (Lijam *et al.,* 1997; section 1.4 for review). The phenotypic similarity of *Notch* and *Pecanex* mutants in *Drosophila* and their similar roles in neural cell fate determination and patterning suggest that they may act in related or possibly the same pathway during development. In summary, the *Pecanex* genes have the classic structure of transmembrane receptors (12 to 14 transmembrane domains). They are without known ligands and have an unknown mode of action but are clearly essential for normal cell fate determination and neural patterning. From a neurodevelomental perspective they are clearly worthy of further study. The position of *PCNXL2* approximately 1.5 to 2 Mb from the chromosome 1 breakpoint places the gene outside the immediate region of interest for the chromosome 1 breakpoint, although this remains conceivably within the range of long range position effects and linkage disequilibrium (sections 1.6.2 and 1.6.3). Testing of association between *PCNXL2* and mental illness in genetic studies may be an interesting and logical development for future studies.

10.6 Future directions I: Comparative genomics

Increasingly, comparative genomics is going to be used as a standard approach for the characterisation of small genomic loci of interest and the annotation of whole genomes. Currently, analysis is typically carried out in a pair-wise manner, in which the observations of conservation can be highly informative, revealing previously unidentified genes, exons or regulatory elements. However, such analysis requires the careful choice of a model genome for comparison. Human to *Fugu* is good for the identification of conserved protein coding regions, but limited for promoter analysis and the identification of small regulatory features. Human to mouse can provide insight into both conserved protein coding regions and regulatory sequences

but a background level of sequence conservation results in a considerable signal to noise problem in distinguishing functional conservation from the paucity of accumulated insertions, deletions and random substitutions.

There will be considerable power in detecting evolutionarily conserved sequences through the alignment of multiple homologous genomic sequences from a range of evolutionarily divergent species. For example, to annotate human genomic sequence it could be aligned to another primate, monkey, dog, cow, mouse, wallaby, platypus, chicken, *Xenopus* and *Fugu*. The patterns of conservation over these ranges of evolutionary distances would build up a highly informative picture of the evolutionary constraints on the locus and changes in constraint over the course of evolution. Along these lines, mouse and chicken clones were isolated for the chromosome 1 breakpoint locus, the sequence of which would complement the comparative analysis carried out between the human and *Fugu* genomic sequences.

Currently tools exist for the comparison of multiple pair-wise genomic sequence alignments (Batzoglou *et al.*, 2000; Schwartz *et al.*, 2000). However, to make full use of multiple homologous genomic sequences it is likely that progressive or iterative alignment methods similar to those of Psi-BLAST (Altschul *et al.*, 1997) for protein sequences will be required. The use of pre-defined orthologous sequence relationships to guide alignments was shown to reduce the complexity of comparative genomic alignment and allowed greater sensitivity in the detection of conserved sequences (section 5.4.2). The incorporation of such techniques with multiple sequence alignments is likely to further increase sensitivity and specificity in the detection of evolutionarily conserved sequences.

The comparison of transcription maps rather than genomic sequence alone was found to be particularly fruitful in gaining insight into the evolutionary constraints of the *TRAX* and *DISC1* genes. The *DISC1* exon 11/11a fusion (section 6.5.1) and the consequent effect on alternative splicing (section 6.5.2), one of the most interesting and informative observations from comparative human versus *Fugu* analysis would not have been observed by pairwise genomic sequence alignment.

10.7 Future directions II: The molecular aetiology of mental illness

Three lines of research following on from the work presented in this thesis will define the molecular pathology leading to mental illness susceptibility in the t(1;11) family and provide new insight into the aetiology of mental illness. (1) The functional characterisation of the genes disrupted by and in close proximity to the translocation breakpoints. (2) Animal models of the translocation. (3) Independent validation of the locus by linkage or association studies.

The *DISC1* gene is the strongest positional candidate by virtue of being directly disrupted by the breakpoint. However, its specific functions remains elusive. The expression patterns of *DISC1* are currently poorly defined. In part, the poorly defined expression pattern reflects the inability to detect *DISC1* mRNA by fluorescent *in situ* hybridisation in the mouse (J. Fantes, personal communication). The zebrafish *DISC1* cDNA sequence reported in section 6.4.2 provides a means of investigating *DISC1* expression in an alternative model organism. Zebrafish is particularly well suited to studying expression patters through the course of development.

The modular nature of the coiled coil regions suggest boundaries of the protein that could be individually investigated for interactions with other proteins through yeast-two-hybrid assay. The conserved distinction between N-terminal and C-terminal regions also suggests a meaningful boundary in the protein to dissect regions of specific interaction and function. The yeast-two-hybrid work has already been initiated with preliminary results supporting the conclusions from comparative sequence analysis that DISC1 forms multiple protein – protein interactions through the formation of heteromeric coiled coil interactions (K. Millar, personal communication). Defining the DISC1 interacting proteins will give new insight into the functions of *DISC1* and provide new functional candidates to be tested for a contribution to the genetic aetiology of mental illness. The yeast-two-hybrid investigations would be complemented by generating antibodies to DISC1 for the further validation of yeast-two-hybrid interaction results and to investigate the expression patterns and sub-cellular localisation of the DISC1 protein.

A variety of animal models are likely to be important for the investigation of both gene function and experimental evaluation of the genes as functional candidates for the t(1;11) associated phenotypes. A mouse *DISC1* knockout may provide insight into the functional roles of *DISC1*. Recapitulating the t(1;11) translocation between the orthologous sequences in the mouse genome is technically feasible using Cre-Lox recombination (Yu and Bradley, 2001 for review). Alternatively, the effect of expressing C-terminally truncated forms of DISC1 in 'normal' and DISC1 null backgrounds may be more readily achieved and informative. Mouse model could provide insight into the function of *DISC1* but may also influence changes in gene expression at other genes such as *TRAX* or *EGLN1*, more faithfully modelling the effects of the translocation in humans. The model mice would be screened both for physical dysmorophism and behavioural abnormalities including tests for prepulse inhibition and other proposed phenotypic markers for schizophrenia susceptibility (section 1.2.4). Brain morphology would also demand particular attention in light of reported brain structural abnormalities in schizophrenics versus controls (section 1.2.3). Potential early developmental roles for *DISC1* could also be investigated using a zebrafish morpholino knockdown strategy (Wixon, 2000 for review). The clone and sequence resources generated through the course of this work represent essential prerequisites to the generation of the animal models described.

The use of a single large family that segregates the phenotype of major mental illness has many advantages for overcoming the problems of phenotypic and genetic heterogeneity (section 1.6.4). The co-segregation of a discrete chromosomal aberration with disease in such a family provides an exceptional opportunity to define the molecular basis of disease in that family. However, independent validation of the molecular lesion is required to demonstrate causation. Support could be derived from gene functional studies and animal models, but validation would require the identification of distinct molecular lesions in the same gene associated with mental illness. The identification of non-translocation families showing significant linkage to the chromosome 1 breakpoint (Hovatta *et al.,* 1998; Ekelund *et al.,* 2001; Gurling *et al.,* 2001; section 1.6.4 for review) provides a

means for genetic validation of the locus. Mutation screening in these families will be an important step in the identification of functional variations. In particular the coding sequences of *DISC1*, *TRAX* and *EGLN1* should be principally targeted for mutation screening based on both functional and positional candidacy. The identified conserved non-coding sequences in the *DISC1 – TRAX – EGLN1* genomic region as well as the *PCNXL2* gene should also be targeted for mutational screening in these families. Subsequently, case - control association studies of each gene are warranted.

In conclusion, since the t(1;11) family and the genetic association with psychiatric illness was first described by St Clair *et al.,* in 1990 the strength of association with the family has increased (section 1.6) and supportive evidence from independent linkage studies have accumulated (section 1.6.4). With these families as a genetic resource as well as contiguous human genomic sequence (section 4.3) and advances in mutation screening technology, there is good reason to predict that a susceptibility locus for major mental illness can be narrowed to a single gene. This thesis has identified and presented many of the candidates and provided new insight into the biology and evolution of the locus.

Reference List

Abrams,R. and Taylor,M.A. (1983). The genetics of schizophrenia: a reassessment using modern criteria. Am. J. Psychiatry *14* , 171-175. PMID: 6849428

Adam,G.I., Reneland,R., Andersson,M., Risinger,C., Nilsson,M., and Lewander,T. (2000). Pharmacogenomics to predict drug response. Pharmacogenomics. *1*, 5-14. PMID: 11258597

Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F., George,R.A., Lewis,S.E., Richards,S., Ashburner,M., and et al. (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195. PMID: 10731132

Akbarian,S., Kim,J.J., Potkin,S.G., Hetrick,W.P., Bunney,W.E., Jr., and Jones,E.G. (1996). Maldistribution of interstitial neurons in prefrontal white matter of the brains of schizophrenic patients. Arch. Gen. Psychiatry *53*, 425-436. PMID: 8624186

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., and Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389-3402. PMID: 9254694

American Psychiatric Association. (1994). Diagnostic and statistical manual fo mental disorders (4th Edn.). American Psychiatric Assoication.

Anchan,R.M. and Reh,T.A. (1995). Transforming growth factor-beta-3 is mitogenic for rat retinal progenitor cells in vitro. J. Neurobiol. *28*, 133-145. PMID: 8537820

Angrist,B., Sathananthan,G., Wilk,S., and Gershon,S. (1974). Amphetamine psychosis: behavioral and biochemical aspects. J. Psychiatr. Res. *11*, 13-23. PMID: 4461784

Aoki,K., Suzuki,K., Sugano,T., Tasaka,T., Nakahara,K., Kuge,O., Omori,A., and Kasai,M. (1995). A novel gene, Translin, encodes a recombination hotspot binding protein associated with chromosomal translocations. Nat. Genet. *10*, 167-174. PMID: 7663511

Aoki,K., Ishida,R., and Kasai,M. (1997). Isolation and characterization of a cDNA encoding a Translin-like protein, TRAX. FEBS Lett. *401*, 109-112. PMID: 9013868

Aoki,K., Inazawa,J., Takahashi,T., Nakahara,K., and Kasai,M. (1997). Genomic structure and chromosomal localization of the gene encoding translin, a recombination hotspot binding protein. Genomics *43*, 237-241. PMID: 9244443

Aoki,K., Suzuki,K., Ishida,R., and Kasai,M. (1999). The DNA binding activity of Translin is mediated by a basic region in the ring-shaped structure conserved in evolution. FEBS Lett. *443*, 363-366. PMID: 10025964

Aparicio,S., Hawker,K., Cottage,A., Mikawa,Y., Zuo,L., Venkatesh,B., Chen,E., Krumlauf,R., and Brenner,S. (1997). Organization of the Fugu rubripes Hox clusters: evidence for continuing evolution of vertebrate Hox complexes. Nat. Genet. *16*, 79-83. PMID: 9140399

Aravind,L. and Koonin,E.V. (2001). The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxogluta. Genome Biol. *2*, RESEARCH0007. PMID: 11276424

Armes,N., Gilley,J., and Fried,M. (1997). The comparative genomic structure and sequence of the surfeit gene homologs in the puffer fish Fugu rubripes and their association with CpG-rich islands. Genome Res. *7*, 1138-1152. PMID: 9414319

Arnold,S.E., Franz,B.R., Gur,R.C., Gur,R.E., Shapiro,R.M., Moberg,P.J., and Trojanowski,J.Q. (1995). Smaller neuron size in schizophrenia in hippocampal subfields that mediate cortical-hippocampal interactions. Am. J. Psychiatry *152*, 738-748. PMID: 7726314

Arriaga,J., Cueto,L., Gonzalez,d.C., and Alarcon-Segovia,D. (1979). [Acute myocardial infarction in a young male with scleroderma (author's. Rev. Invest Clin. *31*, 257-262. PMID: 523860

Aschauer,H.N., Fischer,G., Isenberg,K.E., Meszaros,K., Willinger,U., Todd,R.D., Beran,H., Strobl,R., Lang,M., Fuchs,K., and . (1993). No proof of linkage between schizophrenia-related disorders including schizophrenia and chromosome 2q21 region. Eur. Arch. Psychiatry Clin. Neurosci. *243*, 193-198. PMID: 8117764

Atlas,M., Head,D., Behm,F., Schmidt,E., Zeleznik,L., Roe,B.A., Burian,D., and Domer,P.H. (1998). Cloning and sequence analysis of four t(9;11) therapy-related leukemia breakpoints. Leukemia *12*, 1895-1902. PMID: 9844920

Avner,P. and Heard,E. (2001). X-chromosome inactivation: counting, choice and initiation. Nat. Rev. Genet. *2*, 59-67. PMID: 11253071

Awasthi,S., Palmer,R., Castro,M., Mobarak,C.D., and Ruby,S.W. (2001). New roles for the Snp1 and Exo84 proteins in yeast pre-mRNA splicing. J. Biol. Chem. *276*, 31004-31015. PMID: 11425851

Badge,R.M., Yardley,J., Jeffreys,A.J., and Armour,J.A. (2000). Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. Hum. Mol. Genet. *9*, 1239-1244. PMID: 10767349

Baron,M., Endicott,J., and Ott,J. (1990). Genetic linkage in mental illness. Limitations and prospects. Br. J. Psychiatry *157*, 645-655. PMID: 2279201

Barton,D.E., Kwon,B.S., and Francke,U. (1988). Human tyrosinase gene, mapped to chromosome 11 (q14----q21), defines second region of homology with mouse chromosome 7. Genomics *3*, 17-24. PMID: 3146546

Bassing,C.H., Alt,F.W., Hughes,M.M., D'Auteuil,M., Wehrly,T.D., Woodman,B.B., Gartner,F., White,J.M., Davidson,L., and Sleckman,B.P. (2000). Recombination signal sequences restrict chromosomal V(D)J recombination beyond the 12/23 rule. Nature *405*, 583-586. PMID: 10850719

Batzoglou,S., Pachter,L., Mesirov,J.P., Berger,B., and Lander,E.S. (2000). Human and mouse gene structure: comparative analysis and application to exon prediction. Genome Res. *10*, 950-958. PMID: 10899144

Baxendale,S., Abdulla,S., Elgar,G., Buck,D., Berks,M., Micklem,G., Durbin,R., Bates,G., Brenner,S., and Beck,S. (1995). Comparative sequence analysis of the human and pufferfish Huntington's disease genes. Nat. Genet. *10*, 67-76. PMID: 7647794

Beck,K. and Brodsky,B. (1998). Supercoiled protein motifs: the collagen triple-helix and the alpha-helical coiled coil. J. Struct. Biol. *122*, 17-29. PMID: 9724603

Bell,A.C. and Felsenfeld,G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of. Nature *405*, 482-485. PMID: 10839546

Bell,A.C. and Felsenfeld,G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482-485. PMID: 11549224

Benes,F.M., Kwok,E.W., Vincent,S.L., and Todtenkopf,M.S. (1998). A reduction of nonpyramidal cells in sector CA2 of schizophrenics and manic depressives. Biol. Psychiatry *44*, 88-97. PMID: 9646890

Betz,S., Fairman,R., O'Neil,K., Lear,J., and Degrado,W. (1995). Design of two-stranded and three-stranded coiled-coil peptides. Philos. Trans. R. Soc. Lond B Biol. Sci. *348*, 81-88. PMID: 7770490

Bharath,S., Gangadhar,B.N., and Janakiramaiah,N. (2000). P300 in family studies of schizophrenia: review and critique. Int. J. Psychophysiol. *38*, 43-54. PMID: 11027793

Blackwood,D.H., Whalley,L.J., Christie,J.E., Blackburn,I.M., St Clair,D.M., and McInnes,A. (1987). Changes in auditory P3 event-related potential in schizophrenia and depression. Br. J. Psychiatry *150*, 154-160. PMID: 2888501

Blackwood,D.H., Young,A.H., McQueen,J.K., Martin,M.J., Roxborough,H.M., Muir,W.J., St Clair,D.M., and Kean,D.M. (1991). Magnetic resonance imaging in schizophrenia: altered brain morphology associated with P300 abnormalities and eye tracking dysfunction. Biol. Psychiatry *30*, 753-769. PMID: 1751619

Blackwood,D.H., Fordyce,A., Walker,M.T., St Clair,D.M., Porteous,D.J., and Muir,W.J. (2001). Schizophrenia and affective disorders-cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and p300 findings in a family. Am. J. Hum. Genet. *69*, 428-433. PMID: 11443544

Blake,J.A., Eppig,J.T., Richardson,J.E., Bult,C.J., and Kadin,J.A. (2001). The Mouse Genome Database (MGD): integration nexus for the laboratory mouse. Nucleic Acids Res. *29*, 91-94. PMID: 11125058

Blattner,F.R., Plunkett,G., III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B., and Shao,Y. (1997). The complete genome sequence of Escherichia coli K-12. Science *277*, 1453-1474. PMID: 9278503

Bleuler, M. (1950). Zinkin, J. (Trans.). Dementia praecox or the group of schizophrenias. New York: International Universities Press.

Bliss,T.V. (1999). Young receptors make smart mice. Nature *401*, 25-27. PMID: 10485698

Blouin,J.L., Dombroski,B.A., Nath,S.K., Lasseter,V.K., Wolyniec,P.S., Nestadt,G., Thornquist,M., Ullrich,G., McGrath,J., Kasch,L., Lamacz,M., Thomas,M.G., Gehrig,C., Radhakrishna,U., Snyder,S.E., Balk,K.G., Neufeld,K., Swartz,K.L., DeMarchi,N., Papadimitriou,G.N., Dikeos,D.G., Stefanis,C.N., Chakravarti,A., Childs,B., Pulver,A.E., and . (1998). Schizophrenia susceptibility loci on chromosomes 13q32 and 8p21. Nat. Genet. *20*, 70-73. PMID: 9731535

Blumenthal,T. (1995). Trans-splicing and polycistronic transcription in Caenorhabditis elegans. Trends Genet. *11*, 132-136. PMID: 7732590

Bortolin,M.L. and Kiss,T. (1998). Human U19 intron-encoded snoRNA is processed from a long primary. RNA. *4*, 445-454. PMID: 9630250

Bortolin,M.L. and Kiss,T. (1998). Human U19 intron-encoded snoRNA is processed from a long primary transcript that possesses little potential for protein coding. RNA. *4*, 445-454. PMID: 11549222

Boyd,A.C. (1993). Turbo cloning: a fast, efficient method for cloning PCR products and other blunt-ended DNA fragments into plasmids. Nucleic Acids Res. *21*, 817-821. PMID: 8451184

Brannan,C.I., Dees,E.C., Ingram,R.S., and Tilghman,S.M. (1990). The product of the H19 gene may function as an RNA. Mol. Cell Biol. *10*, 28-36. PMID: 1688465

Bray,N.J. and Owen,M.J. (2001). Searching for schizophrenia genes. Trends Mol. Med. *7*, 169-174. PMID: 11286941

Brenner,D.A. and Bloomer,J.R. (1980). The enzymatic defect in variegate prophyria. Studies with human cultured skin fibroblasts. N. Engl. J. Med. *302*, 765-769. PMID: 7354807

Brenner,S., Elgar,G., Sandford,R., Macrae,A., Venkatesh,B., and Aparicio,S. (1993). Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature *366*, 265-268. PMID: 8232585

Bridger,J.M., Kill,I.R., and Lichter,P. (1998). Association of pKi-67 with satellite DNA of the human genome in early G1 cells. Chromosome. Res. *6*, 13-24. PMID: 9510506

Brown,C.J., Hendrich,B.D., Rupert,J.L., Lafreniere,R.G., Xing,Y., Lawrence,J., and Willard,H.F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that. Cell *71*, 527-542. PMID: 1423611

Brown,C.J., Hendrich,B.D., Rupert,J.L., Lafreniere,R.G., Xing,Y., Lawrence,J., and Willard,H.F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527-542. PMID: 11549223

Bryson,G., Whelahan,H.A., and Bell,M. (2001). Memory and executive function impairments in deficit syndrome schizophrenia. Psychiatry Res. *102*, 29-37. PMID: 11368837

Brzustowicz,L.M., Hodgkinson,K.A., Chow,E.W., Honer,W.G., and Bassett,A.S. (2000). Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22. Science *288*, 678-682. PMID: 10784452

Buchsbaum,M.S., Potkin,S., Siegel,B., Bunney,W.E., Lohr,J., Katz,M., Gottschalk,L., Lottenberg,S., Teng,C., and Abel,L. (1992). PET studies of drug interaction with brain regional glucose metabolism. Clin. Neuropharmacol. *15 Suppl 1 Pt A*, 472A-473A. PMID: 1354054

Bullock,W.O., Fernandez,J.M., and Short,J.M. (1987). XL1-Blue: A high efficiency plasmid transforming recA Eschericia coli strain with beta-galactosidase selection. BioTechniques 376-379. PMID: 11574163

Burge,C. and Karlin,S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. *268*, 78-94. PMID: 9149143

Burge,C.B., Padgett,R.A., and Sharp,P.A. (1998). Evolutionary fates and origins of U12-type introns. Mol. Cell *2*, 773-785. PMID: 9885565

Burset,M. and Guigo,R. (1996). Evaluation of gene structure prediction programs. Genomics *34*, 353-367. PMID: 8786136

Burset,M., Seledtsov,I.A., and Solovyev,V.V. (2001). SpliceDB: database of canonical and non-canonical mammalian splice sites. Nucleic Acids Res. *29*, 255-259. PMID: 11125105

Cannon,T.D., Kaprio,J., Lonnqvist,J., Huttunen,M., and Koskenvuo,M. (1998). The genetic epidemiology of schizophrenia in a Finnish twin cohort. A population-based modeling study. Arch. Gen. Psychiatry *55*, 67-74. PMID: 9435762

Cao,Q., Martinez,M., Zhang,J., Sanders,A.R., Badner,J.A., Cravchik,A., Markey,C.J., Beshah,E., Guroff,J.J., Maxwell,M.E., Kazuba,D.M., Whiten,R., Goldin,L.R., Gershon,E.S., and Gejman,P.V. (1997). Suggestive evidence for a schizophrenia susceptibility locus on chromosome 6q and a confirmation in an independent series of pedigrees. Genomics *43*, 1-8. PMID: 9226366

Cardno,A.G., Marshall,E.J., Coid,B., Macdonald,A.M., Ribchester,T.R., Davies,N.J., Venturi,P., Jones,L.A., Lewis,S.W., Sham,P.C., Gottesman,I.I., Farmer,A.E., McGuffin,P., Reveley,A.M., and Murray,R.M. (1999). Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. Arch. Gen. Psychiatry *56*, 162-168. PMID: 10025441

Cardon,L.R. and Bell,J.I. (2001). Association study designs for complex diseases. Nat. Rev. Genet. *2*, 91-99. PMID: 11253062

Carlsson,M. and Svensson,A. (1990). Interfering with glutamatergic neurotransmission by means of NMDA antagonist administration discloses the locomotor stimulatory potential of other transmitter systems. Pharmacol. Biochem. Behav. *36*, 45-50. PMID: 2161545

Castellanos,F.X., Fine,E.J., Kaysen,D., Marsh,W.L., Rapoport,J.L., and Hallett,M. (1996). Sensorimotor gating in boys with Tourette's syndrome and ADHD: preliminary results. Biol. Psychiatry *39*, 33-41. PMID: 8719124

Castro,A., Peter,M., Magnaghi-Jaulin,L., Vigneron,S., Loyaux,D., Lorca,T., and Labbe,J.C. (2000). Part of Xenopus translin is localized in the centrosomes during mitosis. Biochem. Biophys. Res. Commun. *276*, 515-523. PMID: 11027506

Chalk,J.G., Barr,F.G., and Mitchell,C.D. (1997). Translin recognition site sequences flank chromosome translocation breakpoints in alveolar rhabdomyosarcoma cell lines. Oncogene *15*, 1199-1205. PMID: 9294613

Chang,B.P. and Lenzenweger,M.F. (2001). Somatosensory processing in the biological relatives of schizophrenia patients: a signal detection analysis of two-point discrimination. J. Abnorm. Psychol. *110*, 433-442. PMID: 11502086

Chennathukuzhi,V.M., Kurihara,Y., Bray,J.D., and Hecht,N.B. (2001). Trax (translin-associated factor X), a primarily cytoplasmic protein, inhibits the binding of TB-RBP (translin) to RNA. J. Biol. Chem. *276*, 13256-13263. PMID: 11278549

Chu,R.S., Askew,D., and Harding,C.V. (2000). CpG DNA switches on Th1 immunity and modulates antigen-presenting cell function. Curr. Top. Microbiol. Immunol. *247*, 199-210. PMID: 10689789

Cleary,M.A., van Raamsdonk,C.D., Levorse,J., Zheng,B., Bradley,A., and Tilghman,S.M. (2001). Disruption of an imprinted gene cluster by a targeted chromosomal. Nat. Genet. *29*, 78-82. PMID: 11528397

Cokol,M., Nair,R., and Rost,B. (2000). Finding nuclear localization signals. EMBO Rep. *1*, 411-415. PMID: 11258480

Corbett,R., Hartman,H., Kerman,L.L., Woods,A.T., Strupczewski,J.T., Helsley,G.C., Conway,P.C., and Dunn,R.W. (1993). Effects of atypical antipsychotic agents on social behavior in rodents. Pharmacol. Biochem. Behav. *45*, 9-17. PMID: 7685916

Corbett,R., Camacho,F., Woods,A.T., Kerman,L.L., Fishkin,R.J., Brooks,K., and Dunn,R.W. (1995). Antipsychotic agents antagonize non-competitive N-methyl-D-aspartate antagonist-induced behaviors. Psychopharmacology (Berl) *120*, 67-74. PMID: 7480537

Coyle,J.T. (1996). The glutamatergic dysfunction hypothesis for schizophrenia. Harv. Rev. Psychiatry *3*, 241-253. PMID: 9384954

Crawley,J.N. and Paylor,R. (1997). A proposed test battery and constellations of specific behavioral paradigms to investigate the behavioral phenotypes of transgenic and knockout mice. Horm. Behav. *31*, 197-211. PMID: 9213134

Crick,F.H.C. (1953). The packing of alpha-helicies: simple coiled-coils. Acta Crystallogr. 6, 689-697.

Crow,T.J. (1980). Molecular pathology of schizophrenia: more than one disease process? Br. Med. J. *280*, 66-68. PMID: 6101544

Crowe,R.R. and Vieland,V. (1999). Report of the Chromosome 5 Workshop of the Sixth World Congress on Psychiatric Genetics. Am. J. Med. Genet. *88*, 229-232. PMID: 10374736

Cuff,J.A., Clamp,M.E., Siddiqui,A.S., Finlay,M., and Barton,G.J. (1998). JPred: a consensus secondary structure prediction server. Bioinformatics. *14*, 892-893. PMID: 9927721

D'Arcangelo,G., Miao,G.G., Chen,S.C., Soares,H.D., Morgan,J.I., and Curran,T. (1995). A protein related to extracellular matrix proteins deleted in the mouse mutant reeler. Nature *374*, 719-723. PMID: 7715726

Darby,C., Cosma,C.L., Thomas,J.H., and Manoil,C. (1999). Lethal paralysis of Caenorhabditis elegans by Pseudomonas aeruginosa. Proc. Natl. Acad. Sci. U. S. A *96*, 15202-15207. PMID: 10611362

Davidson,H., Taylor,M.S., Doherty,A., Boyd,A.C., and Porteous,D.J. (2000). Genomic sequence analysis of Fugu rubripes CFTR and flanking genes in a 60 kb region conserving synteny with 800 kb of human chromosome 7. Genome Res. *10*, 1194-1203. PMID: 10958637

Davisson,M.T. and Akeson,E.C. (1993). Recombination suppression by heterozygous Robertsonian chromosomes in the mouse. Genetics *133*, 649-667. PMID: 8454207

Depatie,L. and Lal,S. (2001). Apomorphine and the dopamine hypothesis of schizophrenia: a dilemma? J. Psychiatry Neurosci. *26*, 203-220. PMID: 11394190

des,P., V, Pinard,J.M., Billuart,P., Vinet,M.C., Koulakoff,A., Carrie,A., Gelot,A., Dupuis,E., Motte,J., Berwald-Netter,Y., Catala,M., Kahn,A., Beldjord,C., and Chelly,J. (1998). A novel CNS gene required for neuronal migration and involved in X-linked. Cell *92*, 51-61. PMID: 9489699

Devon,R.S., Evans,K.L., Maule,J.C., Christie,S., Anderson,S., Brown,J., Shibasaki,Y., Porteous,D.J., and Brookes,A.J. (1997). Novel transcribed sequences neighbouring a translocation breakpoint associated with schizophrenia. Am. J. Med. Genet. *74*, 82-90. PMID: 9034012

Devon,R.S., Taylor,M.S., Millar,J.K., and Porteous,D.J. (2000). Isolation and characterization of the mouse translin-associated protein X (Trax) gene. Mamm. Genome *11*, 395-398. PMID: 10790540

Devon,R.S., Anderson,S., Teague,P.W., Burgess,P., Kipari,T.M., Semple,C.A., Millar,J.K., Muir,W.J., Murray,V., Pelosi,A.J., Blackwood,D.H., and Porteous,D.J. (2001). Identification of polymorphisms within Disrupted in Schizophrenia 1 and Disrupted in Schizophrenia 2, and an investigation of their association with schizophrenia and bipolar affective disorder. Psychiatr. Genet. *11*, 71-78. PMID: 11525420

Deybach,J.C., Puy,H., Robreau,A.M., Lamoril,J., Da,S., V, Grandchamp,B., and Nordmann,Y. (1996). Mutations in the protoporphyrinogen oxidase gene in patients with variegate porphyria. Hum. Mol. Genet. *5*, 407-410. PMID: 8852667

Di Cristofano,A., Strazullo,M., Longo,L., and La Mantia,G. (1995). Characterization and genomic mapping of the ZNF80 locus: expression of this zinc-finger gene is driven by a solitary LTR of ERV9 endogenous retroviral family. Nucleic Acids Res. *23*, 2823-2830. PMID: 7659503

Distel,B., Erdmann,R., Gould,S.J., Blobel,G., Crane,D.I., Cregg,J.M., Dodt,G., Fujiki,Y., Goodman,J.M., Just,W.W., Kiel,J.A., Kunau,W.H., Lazarow,P.B., Mannaerts,G.P., Moser,H.W., Osumi,T., Rachubinski,R.A., Roscher,A., Subramani,S., Tabak,H.F., Tsukamoto,T., Valle,D., van,d.K., I, van Veldhoven,P.P., and Veenhuis,M. (1996). A unified nomenclature for peroxisome biogenesis factors. J. Cell Biol. *135*, 1-3. PMID: 8858157

Dobzhansky,T., Sturtevant,A.H. (1931). Translocations between the second and third chromosomes of Drosophila and their baring on Oenothera prolems. Carnegie Institute of Washington Publications, *421*, 29-59.

Dorn,R., Reuter,G., and Loewendorf,A. (2001). Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in Drosophila. Proc. Natl. Acad. Sci. U. S. A *98*, 9724-9729. PMID: 11493677

Dorn,R., Reuter,G., and Loewendorf,A. (2001). Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4). Proc. Natl. Acad. Sci. U. S. A *98*, 9724-9729. PMID: 11528398

Douglas,J.A., Boehnke,M., Gillanders,E., Trent,J.M., and Gruber,S.B. (2001). Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. Nat. Genet. *28*, 361-364. PMID: 11443299

Dupuy,D., Aubert,I., Duperat,V.G., Petit,J., Taine,L., Stef,M., Bloch,B., and Arveiler,B. (2000). Mapping, characterization, and expression analysis of the SM-20 human homologue, c1orf12, and identification of a novel related gene, SCAND2. Genomics *69*, 348-354. PMID: 11056053

Eaton,W.W. (1991). Update on the epidemiology of schizophrenia. Epidemiol. Rev. *13*, 320-328. PMID: 1765116

Edenberg,H.J., Foroud,T., Conneally,P.M., Sorbel,J.J., Carr,K., Crose,C., Willig,C., Zhao,J., Miller,M., Bowman,E., Mayeda,A., Rau,N.L., Smiley,C., Rice,J.P., Goate,A., Reich,T., Stine,O.C., McMahon,F., DePaulo,J.R., Meyers,D., Detera-Wadleigh,S.D., Goldin,L.R., Gershon,E.S., Blehar,M.C., and Nurnberger,J.I., Jr. (1997). Initial genomic scan of the NIMH genetics initiative bipolar pedigrees: chromosomes 3, 5, 15, 16, 17, and 22. Am. J. Med. Genet. *74*, 238-246. PMID: 9184305

Eichler,E.E. (2001). Segmental duplications: what's missing, misassigned, and misassembled--and should we care? Genome Res. *11*, 653-656. PMID: 11337463

Ekelund,J., Hovatta,I., Parker,A., Paunio,T., Varilo,T., Martin,R., Suhonen,J., Ellonen,P., Chan,G., Sinsheimer,J.S., Sobel,E., Juvonen,H., Arajarvi,R., Partonen,T., Suvisaari,J., Lonnqvist,J., Meyer,J., and Peltonen,L. (2001). Chromosome 1 loci in Finnish schizophrenia families. Hum. Mol. Genet. *10*, 1611-1617. PMID: 11468279

Ekker,S.C. (2000). Morphants: a new systematic vertebrate functional genomics approach. Yeast *17*, 302-306. PMID: 11119307

Elgar,G. (1996). Quality not quantity: the pufferfish genome. Hum. Mol. Genet. *5 Spec No*, 1437-1442. PMID: 8875249

Elias,E.R., Mobassaleh,M., Hajra,A.K., and Moser,A.B. (1998). Developmental delay and growth failure caused by a peroxisomal disorder, dihydroxyacetonephosphate acyltransferase (DHAP-AT) deficiency. Am. J. Med. Genet. *80*, 223-226. PMID: 9843043

Eppig,J.T., Blake,J.A., Davisson,M.T., Kadin,J.A., and Richardson,J.E. (2000). The mouse genome database: a resource for today and tomorrow. Lab Anim (NY) *29*, 39-43. PMID: 11375645

Erdmann,V.A., Szymanski,M., Hochberg,A., de Groot,N., and Barciszewski,J. (1999). Collection of mRNA-like non-coding RNAs. Nucleic Acids Res. *27*, 192-195. PMID: 9847177

Erdmann,V.A., Szymanski,M., Hochberg,A., Groot,N., and Barciszewski,J. (2000). Non-coding, mRNA-like RNAs database Y2K. Nucleic Acids Res. *28*, 197-200. PMID: 10592224

Evans,K.L., van,H., V, and Porteous,D.J. (1995). Placement and refined mapping of established and new markers on human chromosome 11q using a small panel of somatic cell hybrids. Eur. J. Hum. Genet. *3*, 42-48. PMID: 7767655

Ewing,B. and Green,P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. *8* , 186-194. PMID: 9521922

Ewing,B., Hillier,L., Wendl,M.C., and Green,P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. *8*, 175-185. PMID: 9521921

Falkai,P., Schneider-Axmann,T., and Honer,W.G. (2000). Entorhinal cortex pre-alpha cell clusters in schizophrenia: quantitative evidence of a developmental abnormality. Biol. Psychiatry *47*, 937-943. PMID: 10838061

Farkas,T., Wolf,A.P., Jaeger,J., Brodie,J.D., Christman,D.R., and Fowler,J.S. (1984). Regional brain glucose metabolism in chronic schizophrenia. A positron emission transaxial tomographic study. Arch. Gen. Psychiatry *41*, 293-300. PMID: 6608333

Feinberg,A.P. and Vogelstein,B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. *132*, 6-13. PMID: 6312838

Feinberg,A.P. and Vogelstein,B. (1984). "A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity". Addendum. Anal. Biochem. *137*, 266-267. PMID: 6329026

Feuchter-Murthy,A.E., Freeman,J.D., and Mager,D.L. (1993). Splicing of a human endogenous retrovirus to a novel phospholipase A2 related gene. Nucleic Acids Res. *21*, 135-143. PMID: 8382789

Finkenstadt,P.M., Kang,W.S., Jeon,M., Taira,E., Tang,W., and Baraban,J.M. (2000). Somatodendritic localization of Translin, a component of the Translin/Trax RNA binding complex. J. Neurochem. *75*, 1754-1762. PMID: 10987859

Finkenstadt,P.M., Jeon,M., and Baraban,J.M. (2001). Masking of the Translin/Trax complex by endogenous RNA. FEBS Lett. *498*, 6-10. PMID: 11389888

Fischer,M., Harvald,B., and Hauge,M. (1969). A Danish twin study of schizophrenia. Br. J. Psychiatry *115*, 981-990. PMID: 5387002

Fletcher,J.M., Evans,K., Baillie,D., Byrd,P., Hanratty,D., Leach,S., Julier,C., Gosden,J.R., Muir,W., Porteous,D.J., and . (1993). Schizophrenia-associated chromosome 11q21 translocation: identification of flanking markers and development of chromosome 11q fragment hybrids as cloning and mapping resources. Am. J. Hum. Genet. *52*, 478-490. PMID: 8383424

Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M., and Miller,W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. *8*, 967-974. PMID: 9750195

Foca,C., Rice,G.E., Quinn,M.A., and Moses,E.K. (2000). Identification and partial characterization of differentially expressed mRNAs in normal human endometria and endometrial carcinomas by differential display RT-PCR. Mol. Hum. Reprod. *6*, 712-718. PMID: 10908281

Ford,J.M. (1999). Schizophrenia: the broken P300 and beyond. Psychophysiology *36*, 667-682. PMID: 10554581

Forrest,D., Yuzaki,M., Soares,H.D., Ng,L., Luk,D.C., Sheng,M., Stewart,C.L., Morgan,J.I., Connor,J.A., and Curran,T. (1994). Targeted disruption of NMDA receptor 1 gene abolishes NMDA response and results in neonatal death. Neuron *13*, 325-338. PMID: 8060614

Franzek,E. and Beckmann,H. (1998). Different genetic background of schizophrenia spectrum psychoses: a twin study. Am. J. Psychiatry *155*, 76-83. PMID: 9433342

Freedman,R., Coon,H., Myles-Worsley,M., Orr-Urtreger,A., Olincy,A., Davis,A., Polymeropoulos,M., Holik,J., Hopkins,J., Hoff,M., Rosenthal,J., Waldo,M.C., Reimherr,F., Wender,P., Yaw,J., Young,D.A., Breese,C.R., Adams,C., Patterson,D., Adler,L.E., Kruglyak,L., Leonard,S., and Byerley,W. (1997). Linkage of a neurophysiological deficit in schizophrenia to a chromosome 15 locus. Proc. Natl. Acad. Sci. U. S. A *94*, 587-592. PMID: 9012828

Futreal,P.A., Kasprzyk,A., Birney,E., Mullikin,J.C., Wooster,R., and Stratton,M.R. (2001). Cancer and genomics. Nature *409*, 850-852. PMID: 11237008

Garver,D.L. (1997). The etiologic heterogeneity of schizophrenia. Harv. Rev. Psychiatry *4*, 317-327. PMID: 9385009

Genest,P., Dumas,L., and Genest,F.B. (1976). [Chromosomal translocation t(2;18)(q21;q23) in a schizophrenic individual and his daughter]. Union Med. Can. *105*, 1676-1681. PMID: 982688

Geyer,M.A., Krebs-Thomson,K., Braff,D.L., and Swerdlow,N.R. (2001). Pharmacological studies of prepulse inhibition models of sensorimotor gating deficits in schizophrenia: a decade in review. Psychopharmacology (Berl) *156*, 117-154. PMID: 11549216

Gilley,J., Armes,N., and Fried,M. (1997). Fugu genome is not a good mammalian model. Nature *385*, 305-306. PMID: 9002512

Gilley,J. and Fried,M. (1999). Extensive gene order differences within regions of conserved synteny between the Fugu and human genomes: implications for chromosomal evolution and the cloning of disease genes. Hum. Mol. Genet. *8*, 1313-1320. PMID: 10369878

Gordon,D., Abajian,C., and Green,P. (1998). Consed: a graphical tool for sequence finishing. Genome Res. *8*, 195-202. PMID: 9521923

Grandy,D.K., Marchionni,M.A., Makam,H., Stofko,R.E., Alfano,M., Frothingham,L., Fischer,J.B., Burke-Howie,K.J., Bunzow,J.R., Server,A.C., and . (1989). Cloning of the cDNA and gene for a human D2 dopamine receptor. Proc. Natl. Acad. Sci. U. S. A *86*, 9762-9766. PMID: 2532362

Granholm,E., Asarnow,R.F., and Marder,S.R. (1991). Controlled information processing resources and the development of automatic detection responses in schizophrenia. J. Abnorm. Psychol. *100*, 22-30. PMID: 2005267

Gratacos,M., Nadal,M., Martin-Santos,R., Pujana,M.A., Gago,J., Peral,B., Armengol,L., Ponsa,I., Miro,R., Bulbena,A., and Estivill,X. (2001). A polymorphic genomic duplication on human chromosome 15 is a susceptibility factor for panic and phobic disorders. Cell *106*, 367-379. PMID: 11509185

Green,M.F., Kern,R.S., Braff,D.L., and Mintz,J. (2000). Neurocognitive deficits and functional outcome in schizophrenia: are we measuring the "right stuff"? Schizophr. Bull. *26*, 119-136. PMID: 10755673

Green,P. (1997). Against a whole-genome shotgun. Genome Res. *7*, 410-417. PMID: 9149937

Gubin,A.N., Njoroge,J.M., Bouffard,G.G., and Miller,J.L. (1999). Gene expression in proliferating human erythroid cells. Genomics *59*, 168-177. PMID: 10409428

Guigo,R., Agarwal,P., Abril,J.F., Burset,M., and Fickett,J.W. (2000). An assessment of gene prediction accuracy in large DNA sequences. Genome Res. *10*, 1631-1642. PMID: 11042160

Guo,W., Grant,A., and Novick,P. (1999). Exo84p is an exocyst protein essential for secretion. J. Biol. Chem. *274*, 23558-23564. PMID: 10438536

Gurling,H.M., Kalsi,G., Brynjolfson,J., Sigmundsson,T., Sherrington,R., Mankoo,B.S., Read,T., Murphy,P., Blaveri,E., McQuillin,A., Petursson,H., and Curtis,D. (2001). Genomewide genetic linkage analysis confirms the presence of susceptibility loci for schizophrenia, on chromosomes 1q32.2, 5q33.2, and 8p21-22 and provides support for linkage to schizophrenia, on chromosomes 11q23.3-24 and 20q12.1-11.23. Am. J. Hum. Genet. *68*, 661-673. PMID: 11179014

Hakak,Y., Walker,J.R., Li,C., Wong,W.H., Davis,K.L., Buxbaum,J.D., Haroutunian,V., and Fienberg,A.A. (2001). Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. Proc. Natl. Acad. Sci. U. S. A *98*, 4746-4751. PMID: 11296301

Hall,R. (1997). Target Schizophrenia. The Assocaition. of the British. Pharmaceutical. Industry. ; Target series. 1-27. PMID: 11549217

Hall,S.L. and Padgett,R.A. (1996). Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. Science *271*, 1716-1718. PMID: 8596930

Han,J.R., Yiu,G.K., and Hecht,N.B. (1995). Testis/brain RNA-binding protein attaches translationally repressed and transported mRNAs to microtubules. Proc. Natl. Acad. Sci. U. S. A *92*, 9550-9554. PMID: 7568171

Hardison,R.C., Oeltjen,J., and Miller,W. (1997). Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res. *7*, 959-966. PMID: 9331366

Harrison,P.J. (1999). The neuropathology of schizophrenia. A critical review of the data and their interpretation. Brain *122 (Pt 4)*, 593-624. PMID: 10219775

Hasegawa,T. and Isobe,K. (1999). Evidence for the interaction between Translin and GADD34 in mammalian cells. Biochim. Biophys. Acta *1428*, 161-168. PMID: 10434033

Hayward,B.E. and Bonthron,D.T. (2000). An imprinted antisense transcript at the human GNAS1 locus. Hum. Mol. Genet. *9*, 835-841. PMID: 10749992

Hecht,N.B. (2000). Intracellular and intercellular transport of many germ cell mRNAs is mediated by the DNA- and RNA-binding protein, testis-brain-RNA-binding protein (TB-RBP). Mol. Reprod. Dev. *56*, 252-253. PMID: 10824978

Heston,L.L. (1966). Psychiatric disorders in foster home reared children of schizophrenic mothers. Br. J. Psychiatry *112*, 819-825. PMID: 5966555

Hiller,W., Dichtl,G., Hecht,H., Hundt,W., and von Zerssen,D. (1994). Testing the comparability of psychiatric diagnoses in ICD-10 and DSM-III-R. Psychopathology *27*, 19-28. PMID: 7972636

Hoffman,R.E., Quinlan,D.M., Mazure,C.M., and McGlashan,T.M. (2001). Cortical instability and the mechanism of mania: a neural network simulation and perceptual test. Biol. Psychiatry *49*, 500-509. PMID: 11257235

Holland,T. and Gosden,C. (1990). A balanced chromosomal translocation partially co-segregating with psychotic illness in a family. Psychiatry Res. *32*, 1-8. PMID: 2349309

Hollander,M.C., Alamo,I., and Fornace,A.J., Jr. (1996). A novel DNA damage-inducible transcript, gadd7, inhibits cell growth, but. Nucleic Acids Res. *24*, 1589-1593. PMID: 8649973

Hollander,M.C., Alamo,I., and Fornace,A.J., Jr. (1996). A novel DNA damage-inducible transcript, gadd7, inhibits cell growth, but lacks a protein product. Nucleic Acids Res. *24*, 1589-1593. PMID: 11549218

Holzman,P.S., Kringlen,E., Levy,D.L., Proctor,L.R., Haberman,S.J., and Yasillo,N.J. (1977). Abnormal-pursuit eye movements in schizophrenia. Evidence for a genetic indicator. Arch. Gen. Psychiatry *34*, 802-805. PMID: 560179

Hosaka,T., Kanoe,H., Nakayama,T., Murakami,H., Yamamoto,H., Nakamata,T., Tsuboyama,T., Oka,M., Kasai,M., Sasaki,M.S., Nakamura,T., and Toguchida,J. (2000). Translin binds to the sequences adjacent to the breakpoints of the TLS and CHOP genes in liposarcomas with translocation t(12;6). Oncogene *19*, 5821-5825. PMID: 11126370

Hovatta,I., Lichtermann,D., Juvonen,H., Suvisaari,J., Terwilliger,J.D., Arajarvi,R., Kokko-Sahin,M.L., Ekelund,J., Lonnqvist,J., and Peltonen,L. (1998). Linkage analysis of putative schizophrenia gene candidate regions on chromosomes 3p, 5q, 6p, 8p, 20p and 22q in a population-based sampled Finnish family set. Mol. Psychiatry *3*, 452-457. PMID: 9774782

Hurst,H.C. (1995). Transcription factors 1: bZIP proteins. Protein Profile. *2*, 101-168. PMID: 7780801

Hurt,E.C. and Schatz,G. (1987). A cytosolic protein contains a cryptic mitochondrial targeting signal. Nature *325*, 499-503. PMID: 3543689

Hytten,F.E. (1976). Is viviparity the best means of reproduction? Acta Paediatr. Acad. Sci. Hung. *17*, 1-8. PMID: 1032866

Imai,K., Harada,S., Kawanishi,Y., Tachikawa,H., Okubo,T., and Suzuki,T. (2001). The (CTG)n polymorphism in the NOTCH4 gene is not associated with schizophrenia in Japanese individuals. BMC. Psychiatry *1*, 1. PMID: 11407996

Impagnatiello,F., Guidotti,A.R., Pesold,C., Dwivedi,Y., Caruncho,H., Pisu,M.G., Uzunov,D.P., Smalheiser,N.R., Davis,J.M., Pandey,G.N., Pappas,G.D., Tueting,P., Sharma,R.P., and Costa,E. (1998). A decrease of reelin expression as a putative vulnerability factor in schizophrenia. Proc. Natl. Acad. Sci. U. S. A *95*, 15718-15723. PMID: 9861036

Jablensky,A., Sartorius,N., Ernberg,G., Anker,M., Korten,A., Cooper,J.E., Day,R., and Bertelsen,A. (1992). Schizophrenia: manifestations, incidence and course in different cultures. A World Health Organization ten-country study. Psychol. Med. Monogr Suppl *20*, 1-97. PMID: 1565705

Jackson,I.J. (1991). A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Res. *19*, 3795-3798. PMID: 1713664

Jacobs,G.H., Stockwell,P.A., Schrieber,M.J., Tate,W.P., and Brown,C.M. (2000). Transterm: a database of messenger RNA components and signals. Nucleic Acids Res. *28*, 293-295. PMID: 10592251

Jacobs,P.A., Aitken,J., Frackiewicz,A., Law,P., Newton,M.S., and Smith,P.G. (1970). The inheritance of translocations in man: data from families ascertained through a balanced heterozygote. Ann. Hum. Genet. *34*, 119-136. PMID: 5493841

Jareborg,N. and Durbin,R. (2000). Alfresco--a workbench for comparative genomic sequence analysis. Genome Res. *10*, 1148-1157. PMID: 10958633

Jaskiw,G.E., Juliano,D.M., Goldberg,T.E., Hertzman,M., Urow-Hamell,E., and Weinberger,D.R. (1994). Cerebral ventricular enlargement in schizophreniform disorder does not progress. A seven year follow-up study. Schizophr. Res. *14*, 23-28. PMID: 7893618

Javitt,D.C. and Zukin,S.R. (1991). Recent advances in the phencyclidine model of schizophrenia. Am. J. Psychiatry *148*, 1301-1308. PMID: 1654746

Jeste,D.V., del Carmen,R., Lohr,J.B., and Wyatt,R.J. (1985). Did schizophrenia exist before the eighteenth century? Compr. Psychiatry *26*, 493-503. PMID: 3905241

Jeste,D.V., Heaton,S.C., Paulsen,J.S., Ercoli,L., Harris,J., and Heaton,R.K. (1996). Clinical and neuropsychological comparison of psychotic depression with nonpsychotic depression and schizophrenia. Am. J. Psychiatry *153*, 490-496. PMID: 8599396

Johnson,C.D. and Liu,L.X. (2000). Novel antimicrobial targets from combined pathogen and host genetics. Proc. Natl. Acad. Sci. U. S. A *97*, 958-959. PMID: 10655466

Kanoe,H., Nakayama,T., Hosaka,T., Murakami,H., Yamamoto,H., Nakashima,Y., Tsuboyama,T., Nakamura,T., Ron,D., Sasaki,M.S., and Toguchida,J. (1999). Characteristics of genomic breakpoints in TLS-CHOP translocations in liposarcomas suggest the involvement of Translin and topoisomerase II in the process of translocation. Oncogene *18*, 721-729. PMID: 9989822

Kasai,M., Matsuzaki,T., Katayanagi,K., Omori,A., Maziarz,R.T., Strominger,J.L., Aoki,K., and Suzuki,K. (1997). The translin ring specifically recognizes DNA ends at recombination hot spots in the human genome. J. Biol. Chem. *272*, 11402-11407. PMID: 9111049

Kato,T. (2001). Molecular genetics of bipolar disorder. Neurosci. Res. *40*, 105-113. PMID: 11377748

Katsanis,J., Taylor,J., Iacono,W.G., and Hammer,M.A. (2000). Heritability of different measures of smooth pursuit eye tracking dysfunction: a study of normal twins. Psychophysiology *37*, 724-730. PMID: 11117452

Kee,Y., Yoo,J.S., Hazuka,C.D., Peterson,K.E., Hsu,S.C., and Scheller,R.H. (1997). Subunit structure of the mammalian exocyst complex. Proc. Natl. Acad. Sci. U. S. A *94*, 14438-14443. PMID: 9405631

Kelley,R.L. and Kuroda,M.I. (2000). Noncoding RNA genes in dosage compensation and imprinting. Cell *103*, 9-12. PMID: 11051542

Kendell,R.E. (1987). Diagnosis and classification of functional psychoses. Br. Med. Bull. *43*, 499-513. PMID: 3322482

Kety,S.S., Rosenthal,D., Wender,P.H., and Schulsinger,F. (1971). Mental illness in the biological and adoptive families of adpoted schizophrenics. Am. J. Psychiatry *128*, 302-306. PMID: 5570994

Kety,S.S., Rosenthal,D., Wender,P.H., Schulsinger,F., and Jacobsen,B. (1976). Mental illness in the biological and adoptive families of adopted individuals who have become schizophrenic. Behav. Genet. *6*, 219-225. PMID: 973827

Kety,S.S. (1988). Schizophrenic illness in the families of schizophrenic adoptees: findings from the Danish national sample. Schizophr. Bull. *14*, 217-222. PMID: 3201179

Kleinjan,D.J. and van,H., V (1998). Position effect in human genetic disease. Hum. Mol. Genet. *7*, 1611-1618. PMID: 9735382

Kobayashi,S., Takashima,A., and Anzai,K. (1998). The dendritic translocation of translin protein in the form of BC1 RNA protein particles in developing rat hippocampal neurons in primary culture. Biochem. Biophys. Res. Commun. *253*, 448-453. PMID: 9878556

Korf,I., Flicek,P., Duan,D., and Brent,M.R. (2001). Integrating genomic homology into gene structure prediction. Bioinformatics. *17 Suppl 1*, S140-S148. PMID: 11473003

Kowalski,P.E., Freeman,J.D., and Mager,D.L. (1999). Intergenic splicing between a HERV-H endogenous retrovirus and two adjacent human genes. Genomics *57*, 371-379. PMID: 10329003

Kozak,M. (1996). Interpreting cDNA sequences: some insights from studies on translation. Mamm. Genome *7*, 563-574. PMID: 8679005

Kraeplin,E. (1902). Clinical psychiatry: A textbook for students and physicians (A.R. Diefendorf, Trans.). New York: Macmillan.

Kringlen,E. (1966). Schizophrenia in twins. An epidemiological-clinical study. Psychiatry *29*, 172-184. PMID: 5943421

Kringlen,E. (2000). Twin studies in schizophrenia with special emphasis on concordance figures. Am. J. Med. Genet. *97*, 4-11. PMID: 10813799

Kuhl,D. and Skehel,P. (1998). Dendritic localization of mRNAs. Curr. Opin. Neurobiol. *8*, 600-606. PMID: 9811623

Kumar,S., Tamura,I.B., Jakobsen,I.B., and Nei. (2001). MEGA2: Molecular evolutionary genetics analysis software. Arizona State University. Tempe, Arizona, USA.

Kutsuwada,T., Sakimura,K., Manabe,T., Takayama,C., Katakura,N., Kushiya,E., Natsume,R., Watanabe,M., Inoue,Y., Yagi,T., Aizawa,S., Arakawa,M., Takahashi,T., Nakamura,Y., Mori,H., and Mishina,M. (1996). Impairment of suckling response, trigeminal neuronal pattern formation, and hippocampal LTD in NMDA receptor epsilon 2 subunit mutant mice. Neuron *16*, 333-344. PMID: 8789948

Kwon,Y.K. and Hecht,N.B. (1991). Cytoplasmic protein binding to highly conserved sequences in the 3' untranslated region of mouse protamine 2 mRNA, a translationally regulated transcript of male germ cells. Proc. Natl. Acad. Sci. U. S. A *88*, 3584-3588. PMID: 2023906

Kwon,Y.K. and Hecht,N.B. (1993). Binding of a phosphoprotein to the 3' untranslated region of the mouse protamine 2 mRNA temporally represses its translation. Mol. Cell Biol. *13*, 6547-6557. PMID: 8413253

LaBonne,S.G., Sunitha,I., and Mahowald,A.P. (1989). Molecular genetics of pecanex, a maternal-effect neurogenic locus of Drosophila melanogaster that potentially encodes a large transmembrane protein. Dev. Biol. *136*, 1-16. PMID: 2478400

Labrador,M., Mongelard,F., Plata-Rengifo,P., Baxter,E.M., Corces,V.G., and Gerasimova,T.I. (2001). Protein encoding by both DNA strands. Nature *409*, 1000. PMID: 11234000

Lander,E. and Kruglyak,L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. *11*, 241-247. PMID: 7581446

Lanz,R.B., McKenna,N.J., Onate,S.A., Albrecht,U., Wong,J., Tsai,S.Y., Tsai,M.J., and O'Malley,B.W. (1999). A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. Cell *97*, 17-27. PMID: 11549219

Lawrie,S.M. and Abukmeil,S.S. (1998). Brain abnormality in schizophrenia. A systematic and quantitative review of volumetric magnetic resonance imaging studies. Br. J. Psychiatry *172*, 110-120. PMID: 9519062

Lebedev,Y.B., Belonovitch,O.S., Zybrova,N.V., Khil,P.P., Kurdyukov,S.G., Vinogradova,T.V., Hunsmann,G., and Sverdlov,E.D. (2000). Differences in HERV-K LTR insertions in orthologous loci of humans and great apes. Gene *247*, 265-277. PMID: 10773466

Lengauer,C., Kinzler,K.W., and Vogelstein,B. (1998). Genetic instabilities in human cancers. Nature *396*, 643-649. PMID: 9872311

Lennox,W.G., Gibbs,E.L., and Gibbs,F.A. The brain-wave pattern, an hereditary trait evidence from 74 normal pairs of twins. Heredity.

Leonhard,K. (1979). [Hereditary and psychosocial causes of schizophrenia]. Psychiatr. Neurol. Med. Psychol. (Leipz.) *31*, 606-626. PMID: 547294

Levinson,D.F., Mahtani,M.M., Nancarrow,D.J., Brown,D.M., Kruglyak,L., Kirby,A., Hayward,N.K., Crowe,R.R., Andreasen,N.C., Black,D.W., Silverman,J.M., Endicott,J., Sharpe,L., Mohs,R.C., Siever,L.J., Walters,M.K., Lennon,D.P., Jones,H.L., Nertney,D.A., Daly,M.J., Gladis,M., and Mowry,B.J. (1998). Genome scan of schizophrenia. Am. J. Psychiatry *155*, 741-750. PMID: 9619145

Li,J., Witte,D.P., Van Dyke,T., and Askew,D.S. (1997). Expression of the putative proto-oncogene His-1 in normal and neoplastic. Am. J. Pathol. *150*, 1297-1305. PMID: 9094986

Lijam,N., Paylor,R., McDonald,M.P., Crawley,J.N., Deng,C.X., Herrup,K., Stevens,K.E., Maccaferri,G., McBain,C.J., Sussman,D.J., and Wynshaw-Boris,A. (1997). Social interaction and sensorimotor gating abnormalities in mice lacking Dvl1. Cell *90*, 895-905. PMID: 9298901

Lindholm,E., Ekholm,B., Shaw,S., Jalonen,P., Johansson,G., Pettersson,U., Sherrington,R., Adolfsson,R., and Jazin,E. (2001). A schizophrenia-susceptibility locus at 6q25, in one of the world's largest reported pedigrees. Am. J. Hum. Genet. *69*, 96-105. PMID: 11389481

Ling,M.L., Risman,S.S., Klement,J.F., McGraw,N., and McAllister,W.T. (1989). Abortive initiation by bacteriophage T3 and T7 RNA polymerases under conditions of limiting substrate. Nucleic Acids Res. *17*, 1605-1618. PMID: 2646596

Linn,G.S. and Javitt,D.C. (2001). Phencyclidine (PCP)-induced deficits of prepulse inhibition in monkeys. Neuroreport *12*, 117-120. PMID: 11201070

Lipscomb,E.A., Sarmiere,P.D., Crowder,R.J., and Freeman,R.S. (1999). Expression of the SM-20 gene promotes death in nerve growth factor-dependent sympathetic neurons. J. Neurochem. *73*, 429-432. PMID: 10386996

Lipscomb,E.A., Sarmiere,P.D., and Freeman,R.S. (2001). SM-20 is a novel mitochondrial protein that causes caspase-dependent cell death in nerve growth factor-dependent neurons. J. Biol. Chem. *276*, 5085-5092. PMID: 11060309

Liu,A.Y., Torchia,B.S., Migeon,B.R., and Siliciano,R.F. (1997). The human NTT gene: identification of a novel 17-kb noncoding nuclear RNA expressed in activated CD4+ T cells. Genomics *39*, 171-184. PMID: 9027504

Lobato,M.I., Belmonte-de-Abreu,P., Knijnik,D., Teruchkin,B., Ghisolfi,E., and Henriques,A. (2001). Neurodevelopmental risk factors in schizophrenia. Braz. J. Med. Biol. Res. *34*, 155-163. PMID: 11175490

Lopez,A.D. and Murray,C.C. (1998). The global burden of disease, 1990-2020. Nat. Med. *4*, 1241-1243. PMID: 9809543

Lussier,I. and Stip,E. (2001). Memory and attention deficits in drug naive patients with schizophrenia. Schizophr. Res. *48*, 45-55. PMID: 11278153

Lyle,R., Watanabe,D., te,V.D., Lerchner,W., Smrzka,O.W., Wutz,A., Schageman,J., Hahner,L., Davies,C., and Barlow,D.P. (2000). The imprinted antisense RNA at the Igf2r locus overlaps but does not. Nat. Genet. *25*, 19-21. PMID: 10802648

Lyle,R., Watanabe,D., te,V.D., Lerchner,W., Smrzka,O.W., Wutz,A., Schageman,J., Hahner,L., Davies,C., and Barlow,D.P. (2000). The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. Nat. Genet. *25*, 19-21. PMID: 11549220

Ma,Z.W., Pejovic,T., Najfeld,V., Ward,D.C., and Johnson,E.M. (1995). Localization of PURA, the gene encoding the sequence-specific single-stranded-DNA-binding protein Pur alpha, to chromosome band 5q31. Cytogenet. Cell Genet. *71*, 64-67. PMID: 7606931

Madden,S.L., Galella,E.A., Riley,D., Bertelsen,A.H., and Beaudry,G.A. (1996). Induction of cell growth regulatory genes by p53. Cancer Res. *56*, 5384-5390. PMID: 8968090

Martinelli,G., Terragna,C., Amabile,M., Montefusco,V., Testoni,N., Ottaviani,E., de Vivo,A., Mianulli,A., Saglio,G., and Tura,S. (2000). Alu and translisin recognition site sequences flanking translocation sites in a novel type of chimeric bcr-abl transcript suggest a possible general mechanism for bcr-abl breakpoints. Haematologica *85*, 40-46. PMID: 10629590

Martinez,M., Goldin,L.R., Cao,Q., Zhang,J., Sanders,A.R., Nancarrow,D.J., Taylor,J.M., Levinson,D.F., Kirby,A., Crowe,R.R., Andreasen,N.C., Black,D.W., Silverman,J.M., Lennon,D.P., Nertney,D.A., Brown,D.M., Mowry,B.J., Gershon,E.S., and Gejman,P.V. (1999). Follow-up study on a susceptibility locus for schizophrenia on chromosome 6q. Am. J. Med. Genet. *88*, 337-343. PMID: 10402499

Masumoto,N., Esaki,T., and Sirotnak,F.M. (2001). Additional organizational features of the murine gamma-glutamyl hydrolase gene. Two remotely situated exons within the complement C3 gene locus encode an alternate 5' end and proximal ORF under the control of a bidirectional promoter. Gene *268*, 183-194. PMID: 11368914

Mathalon,D.H., Ford,J.M., and Pfefferbaum,A. (2000). Trait and state aspects of P300 amplitude reduction in schizophrenia: a retrospective longitudinal study. Biol. Psychiatry *47*, 434-449. PMID: 10704955

Matthysse,S., Holzman,P.S., and Lange,K. (1986). The genetic transmission of schizophrenia: application of Mendelian latent structure analysis to eye tracking dysfunctions in schizophrenia and affective disorder. J. Psychiatr. Res. *20*, 57-67. PMID: 3712291

Mayor,C., Brudno,M., Schwartz,J.R., Poliakov,A., Rubin,E.M., Frazer,K.A., Pachter,L.S., and Dubchak,I. (2000). VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics. *16*, 1046-1047. PMID: 11159318

McCarley,R.W., Faux,S.F., Shenton,M.E., Nestor,P.G., and Adams,J. (1991). Event-related potentials in schizophrenia: their biological and clinical correlates and a new model of schizophrenic pathophysiology. Schizophr. Res. *4*, 209-231. PMID: 2039762

McGinnis,R.E., Fox,H., Yates,P., Cameron,L.A., Barnes,M.R., Gray,I.C., Spurr,N.K., Hurko,O., and St Clair,D. (2001). Failure to confirm NOTCH4 association with schizophrenia in a large population-based sample from Scotland. Nat. Genet. *28*, 128-129. PMID: 11381258

McGlashan,T.H. and Hoffman,R.E. (2000). Schizophrenia as a disorder of developmentally reduced synaptic connectivity. Arch. Gen. Psychiatry *57*, 637-648. PMID: 10891034

McGue,M. and Gottesman,I.I. (1991). The genetic epidemiology of schizophrenia and the design of linkage studies. Eur. Arch. Psychiatry Clin. Neurosci. *240*, 174-181. PMID: 1827604

McGuffin,P., Sargeant,M., Hetti,G., Tidmarsh,S., Whatley,S., and Marchbanks,R.M. (1990). Exclusion of a schizophrenia susceptibility gene from the chromosome 5q11-q13 region: new data and a reanalysis of previous reports. Am. J. Hum. Genet. *47*, 524-535. PMID: 2393025

McLysaght,A., Enright,A.J., Skrabanek,L., and Wolfe,K.H. (2000). Estimation of synteny conservation and genome compaction between pufferfish (Fugu) and human. Yeast *17*, 22-36. PMID: 10797599

McNeil,T.F. and Cantor-Graae,E. (2000). Neuromotor markers of risk for schizophrenia. Aust. N. Z. J. Psychiatry *34 Suppl*, S86-S90. PMID: 11129320

Medstrand,P. and Mager,D.L. (1998). Human-specific integrations of the HERV-K endogenous retrovirus family. J. Virol. *72*, 9782-9787. PMID: 9811713

Meng,G., Aoki,K., Tokura,K., Nakahara,K., Inazawa,J., and Kasai,M. (2000). Genomic structure and chromosomal localization of the gene encoding TRAX, a Translin-associated factor X. J. Hum. Genet. *45*, 305-308. PMID: 11043515

Michie,P.T., Kent,A., Stienstra,R., Castine,R., Johnston,J., Dedman,K., Wichmann,H., Box,J., Rock,D., Rutherford,E., and Jablensky,A. (2000). Phenotypic markers as risk factors in schizophrenia: neurocognitive functions. Aust. N. Z. J. Psychiatry *34 Suppl*, S74-S85. PMID: 11129319

Milanesi,L., D'Angelo,D., and Rogozin,I.B. (1999). GeneBuilder: interactive in silico prediction of gene structure. Bioinformatics. *15*, 612-621. PMID: 10487869

Miles,C., Elgar,G., Coles,E., Kleinjan,D.J., van,H., V, and Hastie,N. (1998). Complete sequencing of the Fugu WAGR region from WT1 to PAX6: dramatic compaction and conservation of synteny with human chromosome 11p13. Proc. Natl. Acad. Sci. U. S. A *95*, 13068-13072. PMID: 9789042

Millar,J.K., Christie,S., Semple,C.A., and Porteous,D.J. (2000b). Chromosomal location and genomic structure of the human translin-associated factor X gene (TRAX; TSNAX) revealed by intergenic splicing to DISC1, a gene disrupted by a translocation segregating with schizophrenia. Genomics *67*, 69-77. PMID: 10945471

Millar,J.K., Wilson-Annan,J.C., Anderson,S., Christie,S., Taylor,M.S., Semple,C.A., Devon,R.S., Clair,D.M., Muir,W.J., Blackwood,D.H., and Porteous,D.J. (2000a). Disruption of two novel genes by a translocation co-segregating with schizophrenia. Hum. Mol. Genet. *9*, 1415-1423. PMID: 10814723

Millar,J.K., Christie,S., Anderson,S., Lawson,D., Hsiao-Wei,L.D., Devon,R.S., Arveiler,B., Muir,W.J., Blackwood,D.H., and Porteous,D.J. (2001). Genomic structure and localisation within a linkage hotspot of Disrupted In Schizophrenia 1, a gene disrupted by a translocation segregating with schizophrenia. Mol. Psychiatry *6*, 173-178. PMID: 11317219

Miller,S.A., Dykes,D.D., and Polesky,H.F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. *16*, 1215. PMID: 3344216

Mlynarczyk,S.K. and Panning,B. (2000). X inactivation: Tsix and Xist as yin and yang. Curr. Biol. *10*, R899-R903. PMID: 11137025

Mohn,A.R., Gainetdinov,R.R., Caron,M.G., and Koller,B.H. (1999). Mice with reduced NMDA receptor expression display behaviors related to schizophrenia. Cell *98*, 427-436. PMID: 10481908

Moises,H.W., Yang,L., Kristbjarnarson,H., Wiese,C., Byerley,W., Macciardi,F., Arolt,V., Blackwood,D., Liu,X., Sjogren,B., and . (1995). An international two-stage genome-wide search for schizophrenia susceptibility genes. Nat. Genet. *11*, 321-324. PMID: 7581457

Moldin,S.O. (1997). The maddening hunt for madness genes. Nat. Genet. *17*, 127-129. PMID: 9326920

Moore,M.J. (2000). Intron recognition comes of AGe. Nat. Struct. Biol. *7*, 14-16. PMID: 10625417

Morales,C.R., Wu,X.Q., and Hecht,N.B. (1998). The DNA/RNA-binding protein, TB-RBP, moves from the nucleus to the cytoplasm and through intercellular bridges in male germ cells. Dev. Biol. *201*, 113-123. PMID: 9733578

Morissette,J., Villeneuve,A., Bordeleau,L., Rochette,D., Laberge,C., Gagne,B., Laprise,C., Bouchard,G., Plante,M., Gobeil,L., Shink,E., Weissenbach,J., and Barden,N. (1999). Genome-wide search for linkage of bipolar affective disorders in a very large pedigree derived from a homogeneous population in quebec points to a locus of major effect on chromosome 12q23-q24. Am. J. Med. Genet. *88*, 567-587. PMID: 10490718

Moschella,M.C., Menzies,K., Tsao,L., Lieb,M.A., Kohtz,J.D., Kohtz,D.S., and Taubman,M.B. (1999). SM-20 is a novel growth factor-responsive gene regulated during skeletal muscle development and differentiation. Gene Expr. *8*, 59-66. PMID: 10543731

Mott,R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput. Appl. Biosci. *13*, 477-478. PMID: 9283765

Muller,K.M., Arndt,K.M., and Alber,T. (2000). Protein fusions to coiled-coil domains. Methods Enzymol. *328*, 261-282. PMID: 11075350

Muramatsu,T., Ohmae,A., and Anzai,K. (1998). BC1 RNA protein particles in mouse brain contain two y-,h-element-binding proteins, translin and a 37 kDa protein. Biochem. Biophys. Res. Commun. *247*, 7-11. PMID: 9636644

Murray,R.M., Lewis,S.W., and Reveley,A.M. (1985). Towards an aetiological classification of schizophrenia. Lancet *1*, 1023-1026. PMID: 2859472

Myers,E.W., Sutton,G.G., Delcher,A.L., Dew,I.M., Fasulo,D.P., Flanigan,M.J., Kravitz,S.A., Mobarry,C.M., Reinert,K.H., Remington,K.A., Anson,E.L., Bolanos,R.A., Chou,H.H., Jordan,C.M., Halpern,A.L., Lonardi,S., Beasley,E.M., Brandon,R.C., Chen,L., Dunn,P.J., Lai,Z., Liang,Y., Nusskern,D.R., Zhan,M., Zhang,Q., Zheng,X., Rubin,G.M., Adams,M.D., and Venter,J.C. (2000). A whole-genome assembly of Drosophila. Science *287*, 2196-2204. PMID: 10731133

Nakai,K. and Kanehisa,M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics *14* , 897-911. PMID: 1478671

Nakai,K. and Horton,P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem. Sci. *24*, 34-36. PMID: 10087920

Nelson,D.T., Goodchild,N.L., and Mager,D.L. (1996). Gain of Sp1 sites and loss of repressor sequences associated with a young, transcriptionally active subset of HERV-H endogenous long terminal repeats. Virology *220*, 213-218. PMID: 8659116

Ninomiya,S., Isomura,M., Narahara,K., Seino,Y., and Nakamura,Y. (1996). Isolation of a testis-specific cDNA on chromosome 17q from a region. Hum. Mol. Genet. *5*, 69-72. PMID: 8789441

Ninomiya,S., Isomura,M., Narahara,K., Seino,Y., and Nakamura,Y. (1996). Isolation of a testis-specific cDNA on chromosome 17q from a region adjacent to the breakpoint of t(12;17) observed in a patient with acampomelic campomelic dysplasia and sex reversal. Hum. Mol. Genet. *5*, 69-72. PMID: 11549221

O'Donovan,M.C. and Owen,M.J. (1999). Candidate-gene association studies of schizophrenia. Am. J. Hum. Genet. *65*, 587-592. PMID: 10441563

Ohara,K., Xu,H.D., Matsunaga,T., Xu,D.S., Huang,X.Q., Gu,G.F., Ohara,K., and Wang,Z.C. (1998). Cerebral ventricle-brain ratio in monozygotic twins discordant and concordant for schizophrenia. Prog. Neuropsychopharmacol. Biol. Psychiatry *22*, 1043-1050. PMID: 9789887

Ohashi,S., Kobayashi,S., Omori,A., Ohara,S., Omae,A., Muramatsu,T., Li,Y., and Anzai,K. (2000). The single-stranded DNA- and RNA-binding proteins pur alpha and pur beta link BC1 RNA to microtubules through binding to the dendrite-targeting RNA motifs. J. Neurochem. *75*, 1781-1790. PMID: 11032866

Olsen,D.R., Nagayoshi,T., Fazio,M., Mattei,M.G., Passage,E., Weil,D., Timpl,R., Chu,M.L., and Uitto,J. (1989). Human nidogen: cDNA cloning, cellular expression, and mapping of the gene to chromosome Iq43. Am. J. Hum. Genet. *44*, 876-885. PMID: 2471408

Onyango,P., Miller,W., Lehoczky,J., Leung,C.T., Birren,B., Wheelan,S., Dewar,K., and Feinberg,A.P. (2000). Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain. Genome Res. *10*, 1697-1710. PMID: 11076855

Osoegawa,K., Tateno,M., Woon,P.Y., Frengen,E., Mammoser,A.G., Catanese,J.J., Hayashizaki,Y., and de Jong,P.J. (2000). Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. Genome Res. *10*, 116-128. PMID: 10645956

Ott,J. (2001). Major strengths and weaknesses of the lod score method. Adv. Genet. *42*, 125-132. PMID: 11037318

Owen,M., Craufurd,D., and St Clair,D. (1990). Localisation of a susceptibility locus for schizophrenia on chromosome 5. Br. J. Psychiatry *157*, 123-127. PMID: 1975758

Pangalos,M.N., Neefs,J.M., Somers,M., Verhasselt,P., Bekkers,M., van der,H.L., Fraiponts,E., Ashton,D., and Gordon,R.D. (1999). Isolation and expression of novel human glutamate carboxypeptidases with N-acetylated alpha-linked acidic dipeptidase and dipeptidyl peptidase IV activity. J. Biol. Chem. *274*, 8470-8483. PMID: 10085079

Pearson,W.R., Wood,T., Zhang,Z., and Miller,W. (1997). Comparison of DNA sequences with protein sequences. Genomics *46*, 24-36. PMID: 9403055

Pearson,W.R., Robins,G., and Zhang,T. (1999). Generalized neighbor-joining: more reliable phylogenetic tree reconstruction. Mol. Biol. Evol. *16*, 806-816. PMID: 10368958

Pelczar,P. and Filipowicz,W. (1998). The host gene for intronic U17 small nucleolar RNAs in mammals has no protein-coding potential and is a member of the 5'-terminal oligopyrimidine gene family. Mol. Cell Biol. *18*, 4509-4518. PMID: 9671460

Persson,B. and Argos,P. (1994). Prediction of transmembrane segments in proteins utilising multiple sequence alignments. J. Mol. Biol. *237*, 182-192. PMID: 8126732

Pesole,G., Liuni,S., Grillo,G., Licciulli,F., Larizza,A., Makalowski,W., and Saccone,C. (2000). UTRdb and UTRsite: specialized databases of sequences and functional. Nucleic Acids Res. *28*, 193-196. PMID: 10592223

Phan Dinh,T.F., Porteu,A., Kahn,A., and Skala,H. (2001). Bidirectional activity and orientation-dependent specificity of the rat aldolase C promoter in transgenic mice. FEBS Lett. *499*, 143-146. PMID: 11418129

Pittack,C., Grunwald,G.B., and Reh,T.A. (1997). Fibroblast growth factors are necessary for neural retina but not pigmented epithelium differentiation in chick embryos. Development *124*, 805-816. PMID: 9043062

Pope,H.G., Jr., Jonas,J.M., Cohen,B.M., and Lipinski,J.F. (1982). Failure to find evidence of schizophrenia in first-degree relatives of schizophrenic probands. Am. J. Psychiatry *139*, 826-828. PMID: 7081502

Prescott,C.A. and Gottesman,I.I. (1993). Genetically mediated vulnerability to schizophrenia. Psychiatr. Clin. North Am. *16*, 245-267. PMID: 8332563

Pulver,A.E., Mulle,J., Nestadt,G., Swartz,K.L., Blouin,J.L., Dombroski,B., Liang,K.Y., Housman,D.E., Kazazian,H.H., Antonarakis,S.E., Lasseter,V.K., Wolyniec,P.S., Thornquist,M.H., and McGrath,J.A. (2000). Genetic heterogeneity in schizophrenia: stratification of genome scan data using co-segregating related phenotypes. Mol. Psychiatry *5*, 650-653. PMID: 11126395

Reinhardt,A. and Hubbard,T. (1998). Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res. *26*, 2230-2236. PMID: 9547285

Riordan,J.R., Rommens,J.M., Kerem,B., Alon,N., Rozmahel,R., Grzelczak,Z., Zielenski,J., Lok,S., Plavsic,N., Chou,J.L., and . (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science *245*, 1066-1073. PMID: 2475911

Risch,N. and Baron,M. (1984). Segregation analysis of schizophrenia and related disorders. Am. J. Hum. Genet. *36*, 1039-1059. PMID: 6496472

Risch,N. (1990). Genetic linkage and complex diseases, with special reference to psychiatric disorders. Genet. Epidemiol. *7*, 3-16. PMID: 2184091

Risch,N. (1992). Genetic linkage: interpreting lod scores. Science *255*, 803-804. PMID: 1536004

Roach,P.L., Clifton,I.J., Fulop,V., Harlos,K., Barton,G.J., Hajdu,J., Andersson,I., Schofield,C.J., and Baldwin,J.E. (1995). Crystal structure of isopenicillin N synthase is the first from a new structural family of enzymes. Nature *375*, 700-704. PMID: 7791906

Robinson,D., Woerner,M.G., Alvir,J.M., Bilder,R., Goldman,R., Geisler,S., Koreen,A., Sheitman,B., Chakos,M., Mayerhoff,D., and Lieberman,J.A. (1999). Predictors of relapse following response from a first episode of schizophrenia or schizoaffective disorder. Arch. Gen. Psychiatry *56*, 241-247. PMID: 10078501

Rosenthall,D., Wender,P.H., Kety,S.S., Schulsinger,F., Welner,J., and Ostergaard,L. (1968). Schizophrenics' offspring raised in adoptive homes.

Rost,B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol. *266*, 525-539. PMID: 8743704

Rund,B.R. and Borg,N.E. (1999). Cognitive deficits and cognitive training in schizophrenic patients: a review. Acta Psychiatr. Scand. *100*, 85-95. PMID: 10480194

Sambrook,J., Fritsch,E.F., and Maniatis,T. (1989). Molecular cloning: A laboratory manual. CSH Laboratory Press, Cold Spring Harbour.

Sayle,R.A. and Milner-White,E.J. (1995). RASMOL: biomolecular graphics for all. Trends Biochem. Sci. *20*, 374. PMID: 7482707

Scacheri,P.C., Crabtree,J.S., Novotny,E.A., Garrett-Beal,L., Chen,A., Edgemon,K.A., Marx,S.J., Spiegel,A.M., Chandrasekharappa,S.C., and Collins,F.S. (2001). Bidirectional transcriptional activity of PGK-neomycin and unexpected embryonic lethality in heterozygote chimeric knockout mice. Genesis. *30*, 259-263. PMID: 11536432

Scharfetter,J. (2001). Dopamine receptor polymorphisms and drug response in schizophrenia. Pharmacogenomics. *2*, 251-261. PMID: 11535113

Schreiber,H., Stolz-Born,G., Kornhuber,H.H., and Born,J. (1992). Event-related potential correlates of impaired selective attention in. Biol. Psychiatry *32*, 634-651. PMID: 1457620

Schroder,J., Buchsbaum,M.S., Siegel,B.V., Geider,F.J., and Niethammer,R. (1995). Structural and functional correlates of subsyndromes in chronic schizophrenia. Psychopathology *28*, 38-45. PMID: 7871119

Schuler,G.D. (1997). Pieces of the puzzle: expressed sequence tags and the catalog of human genes. J. Mol. Med. *75*, 694-698. PMID: 9382993

Schwab,S.G., Albus,M., Hallmayer,J., Honig,S., Borrmann,M., Lichtermann,D., Ebstein,R.P., Ackenheil,M., Lerer,B., Risch,N. (1995). Evaluation of a susceptibility gene for schizophrenia on chromosome 6p by multipoint affected sib-pair linkage analysis. Nat. Genet. *11*, 325-327. PMID: 7581458

Schwab,S.G., Eckstein,G.N., Hallmayer,J., Lerer,B., Albus,M., Borrmann,M., Lichtermann,D., Ertl,M.A., Maier,W., and Wildenauer,D.B. (1997). Evidence suggestive of a locus on chromosome 5q31 contributing to susceptibility for schizophrenia in German and Israeli families by multipoint affected sib-pair linkage analysis. Mol. Psychiatry *2*, 156-160. PMID: 9106241

Schwartz,S., Zhang,Z., Frazer,K.A., Smit,A., Riemer,C., Bouck,J., Gibbs,R., Hardison,R., and Miller,W. (2000). PipMaker--a web server for aligning two genomic DNA sequences. Genome Res. *10*, 577-586. PMID: 10779500

Selemon,L.D., Rajkowska,G., and Goldman-Rakic,P.S. (1998). Elevated neuronal density in prefrontal area 46 in brains from schizophrenic patients: application of a three-dimensional, stereologic counting method. J. Comp Neurol. *392*, 402-412. PMID: 9511926

Semple,C.A., Devon,R.S., Le Hellard,S., and Porteous,D.J. (2001). Identification of genes from a schizophrenia-linked translocation breakpoint region. Genomics *73*, 123-126. PMID: 11352574

Semple C.A., Morris S.W., Porteous D.J., Evans K.L. (2002). Computational comparison of human genomic sequence assemblies for a region of chromosome 4. Genome Res. *12*, 424-429. PMID: 11875030

Semple,C.A., Taylor,M.S., and Ballereau,S. (2001). The meso-genomic era. Genome Biol. *2*, REPORTS4015. PMID: 11516332

Sengoku,A. and Takagi,S. (1998). Electroencephalographic findings in functional psychoses: state or trait indicators? Psychiatry Clin. Neurosci. *52*, 375-381. PMID: 9766684

Seoighe,C., Federspiel,N., Jones,T., Hansen,N., Bivolarovic,V., Surzycki,R., Tamse,R., Komp,C., Huizar,L., Davis,R.W., Scherer,S., Tait,E., Shaw,D.J., Harris,D., Murphy,L., Oliver,K., Taylor,K.,

Rajandream,M.A., Barrell,B.G., and Wolfe,K.H. (2000). Prevalence of small inversions in yeast gene order evolution. Proc. Natl. Acad. Sci. U. S. A *97*, 14433-14437. PMID: 11087826

Severt,W.L., Biber,T.U., Wu,X., Hecht,N.B., DeLorenzo,R.J., and Jakoi,E.R. (1999). The suppression of testis-brain RNA binding protein and kinesin heavy chain disrupts mRNA sorting in dendrites. J. Cell Sci. *112 (Pt 21)*, 3691-3702. PMID: 10523505

Sham,P.C., Morton,N.E., Muir,W.J., Walker,M., Collins,A., Shields,D.C., St Clair,D.M., and Blackwood,D.H. (1994). Segregation analysis of complex phenotypes: an application to. Psychiatr. Genet. *4*, 29-38. PMID: 8049901

Sharp,P.A. and Burge,C.B. (1997). Classification of introns: U2-type or U12-type. Cell *91*, 875-879. PMID: 9428511

Sherrington,R., Brynjolfsson,J., Petursson,H., Potter,M., Dudleston,K., Barraclough,B., Wasmuth,J., Dobbs,M., and Gurling,H. (1988). Localization of a susceptibility locus for schizophrenia on chromosome 5. Nature *336*, 164-167. PMID: 2903449

Silva,A.J., Elgersma,Y., and Costa,R.M. (2000). Molecular and cellular mechanisms of cognitive function: implications for psychiatric disorders. Biol. Psychiatry *47*, 200-209. PMID: 10682217

Siris,S.G. (2001). Suicide and schizophrenia. J. Psychopharmacol. *15*, 127-135. PMID: 11448086

Sklar,P., Schwab,S.G., Williams,N.M., Daly,M., Schaffner,S., Maier,W., Albus,M., Trixler,M., Eichhammer,P., Lerer,B., Hallmayer,J., Norton,N., Williams,H., Zammit,S., Cardno,A.G., Jones,S., McCarthy,G., Milanova,V., Kirov,G., O'Donovan,M.C., Lander,E.S., Owen,M.J., and Wildenauer,D.B. (2001). Association analysis of NOTCH4 loci in schizophrenia using family and population-based controls. Nat. Genet. *28*, 126-128. PMID: 11381257

Smith,C.M. and Steitz,J.A. (1998). Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. Mol. Cell Biol. *18*, 6897-6909. PMID: 9819378

Smith,M., Wasmuth,W., McPherson,J.D., Wagner,C., Grandy,D., Civelli,O., Potkin,S., and Litt,M. (1989). Cosegregation of an 1q22-9p22 translocation with affective disorder: proximity to the dopamine D2 receptor relative to the translocation breakpoint. A220. PMID: 11574162

Sonnhammer,E.L. and Durbin,R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene *167*, GC1-10. PMID: 8566757

Southern,E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. *98* , 503-517. PMID: 1195397

Spitzer,R.L., Endicott,J., and Robins,E. (1978). Research diagnostic criteria: rationale and reliability. Arch. Gen. Psychiatry *35*, 773-782. PMID: 655775

St Clair,D., Blackwood,D., Muir,W., Carothers,A., Walker,M., Spowart,G., Gosden,C., and Evans,H.J. (1990). Association within a family of a balanced autosomal translocation with major mental illness. Lancet *336*, 13-16. PMID: 11574161

St Clair,D., Blackwood,D., Muir,W., Carothers,A., Walker,M., Spowart,G., Gosden,C., and Evans,H.J. (1990). Association within a family of a balanced autosomal translocation with major mental illness. Lancet *336*, 13-16. PMID: 1973210

Stein, L. How Perl saved the human genome project. The Perl Journal . 2-2-1996. PMID: 11493678

Stober,G., Saar,K., Ruschendorf,F., Meyer,J., Nurnberg,G., Jatzke,S., Franzek,E., Reis,A., Lesch,K.P., Wienker,T.F., and Beckmann,H. (2000). Splitting schizophrenia: periodic catatonia-susceptibility locus on chromosome 15q15. Am. J. Hum. Genet. *67*, 1201-1207. PMID: 11001582

Straub,R.E., MacLean,C.J., O'Neill,F.A., Burke,J., Murphy,B., Duke,F., Shinkwin,R., Webb,B.T., Zhang,J., Walsh,D., and . (1995). A potential vulnerability locus for schizophrenia on chromosome 6p24-22: evidence for genetic heterogeneity. Nat. Genet. *11*, 287-293. PMID: 7581452

Straub,R.E., MacLean,C.J., O'Neill,F.A., Walsh,D., and Kendler,K.S. (1997). Support for a possible schizophrenia vulnerability locus in region 5q22-31 in Irish families. Mol. Psychiatry *2*, 148-155. PMID: 9106240

Suddath,R.L., Christison,G.W., Torrey,E.F., Casanova,M.F., and Weinberger,D.R. (1990). Anatomical abnormalities in the brains of monozygotic twins discordant for schizophrenia. N. Engl. J. Med. *322*, 789-794. PMID: 2308615

Sutherland,H.F., Wadey,R., McKie,J.M., Taylor,C., Atif,U., Johnstone,K.A., Halford,S., Kim,U.J., Goodship,J., Baldini,A., and Scambler,P.J. (1996). Identification of a novel transcript disrupted by a balanced translocation associated with DiGeorge syndrome. Am. J. Hum. Genet. *59*, 23-31. PMID: 8659529

Sverdlov,E.D. (2000). Retroviruses and primate evolution. Bioessays *22*, 161-171. PMID: 10655035

Swerdlow,N.R., Benbow,C.H., Zisook,S., Geyer,M.A., and Braff,D.L. (1993). A preliminary assessment of sensorimotor gating in patients with obsessive compulsive disorder. Biol. Psychiatry *33*, 298-301. PMID: 8471686

Swerdlow,N.R., Paulsen,J., Braff,D.L., Butters,N., Geyer,M.A., and Swenson,M.R. (1995). Impaired prepulse inhibition of acoustic and tactile startle response in patients with Huntington's disease. J. Neurol. Neurosurg. Psychiatry *58*, 192-200. PMID: 7876851

Swerdlow,N.R., Braff,D.L., and Geyer,M.A. (1999). Cross-species studies of sensorimotor gating of the startle reflex. Ann. N. Y. Acad. Sci. *877*, 202-216. PMID: 10415651

Taira,E., Finkenstadt,P.M., and Baraban,J.M. (1998). Identification of translin and trax as components of the GS1 strand-specific DNA binding complex enriched in brain. J. Neurochem. *71*, 471-477. PMID: 9681436

Tamminga,C.A. (1998). Schizophrenia and glutamatergic transmission. Crit Rev. Neurobiol. *12*, 21-36. PMID: 9444480

Tang,Y.P., Shimizu,E., Dube,G.R., Rampon,C., Kerchner,G.A., Zhuo,M., Liu,G., and Tsien,J.Z. (1999). Genetic enhancement of learning and memory in mice. Nature *401*, 63-69. PMID: 10485705

Tarn,W.Y. and Steitz,J.A. (1996). Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. Science *273*, 1824-1832. PMID: 8791582

Taylor,M.S. (2001a). More biology from the sequence. Genome Biol. *2*, REPORTS4018. PMID: 11532209

Taylor,M.S. (2001b). Characterization and comparative analysis of the EGLN gene family. Gene *275*, 125-132. PMID: 11574160

Terwilliger,J.D. and Ott,J. (1992). A multisample bootstrap approach to the estimation of maximized-over-models lod score distributions. Cytogenet. Cell Genet. *59*, 142-144. PMID: 1737483

Thai,T.P., Heid,H., Rackwitz,H.R., Hunziker,A., Gorgas,K., and Just,W.W. (1997). Ether lipid biosynthesis: isolation and molecular characterization of human dihydroxyacetonephosphate acyltransferase. FEBS Lett. *420*, 205-211. PMID: 9459311

Thaker,G.K. (2000). Defining the schizophrenia phenotype. Curr. Psychiatry Rep. *2*, 398-403. PMID: 11122987

Thaker,G.K. and Carpenter,W.T., Jr. (2001). Advances in schizophrenia. Nat. Med. *7*, 667-671. PMID: 11385502

The international human genome sequencing consortium (2001). Initial sequencing and analysis of the human genome. Nature *409* , 860-921. PMID: 11237011

Thiery,J.P., Macaya,G., and Bernardi,G. (1976). An analysis of eukaryotic genomes by density gradient centrifugation. J. Mol. Biol. *108*, 219-235. PMID: 826643

Thompson,J.D., Higgins,D.G., and Gibson,T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673-4680. PMID: 7984417

Tienari,P. (1963). Psychiatric illness in identical twins. Acta Psychiatr. Scand.

Tienari,P., Wynne,L.C., Moring,J., Lahti,I., Naarala,M., Sorri,A., Wahlberg,K.E., Saarento,O., Seitamaa,M., Kaleva,M., and . (1994). The Finnish adoptive family study of schizophrenia. Implications for family research. Br. J. Psychiatry Suppl 20-26. PMID: 8037897

Ting,Y.M., Doust,B.D., and Chuang,V.P. (1975). Xerotomographic diagnosis of central bronchogenic carcinoma. Chest *67*, 172-175. PMID: 1116392

Valegard,K., van Scheltinga,A.C., Lloyd,M.D., Hara,T., Ramaswamy,S., Perrakis,A., Thompson,A., Lee,H.J., Baldwin,J.E., Schofield,C.J., Hajdu,J., and Andersson,I. (1998). Structure of a cephalosporin synthase. Nature *394*, 805-809. PMID: 9723623

van Eeden,F. and St Johnston,D. (1999). The polarisation of the anterior-posterior and dorsal-ventral axes during Drosophila oogenesis. Curr. Opin. Genet. Dev. *9*, 396-404. PMID: 10449356

Velleca,M.A., Wallace,M.C., and Merlie,J.P. (1994). A novel synapse-associated noncoding RNA. Mol. Cell Biol. *14*, 7095-7104. PMID: 7523860

Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., and et al. (2001). The sequence of the human genome. Science *291*, 1304-1351. PMID: 11181995

Vervoort,R., Lennon,A., Bird,A.C., Tulloch,B., Axton,R., Miano,M.G., Meindl,A., Meitinger,T., Ciccodicola,A., and Wright,A.F. (2000). Mutational hot spot within a new RPGR exon in X-linked retinitis pigmentosa. Nat. Genet. *25*, 462-466. PMID: 10932196

Vogel,F., Schalt,E., and Kruger,J. (1979). The electroencephalogram (EEG) as a research tool in human behavior genetics: psychological examinations in healthy males with various inherited EEG variants. II. Results. Hum. Genet. *47*, 47-80. PMID: 429013

Wang,S., Sun,C.E., Walczak,C.A., Ziegle,J.S., Kipps,B.R., Goldin,L.R., and Diehl,S.R. (1995). Evidence for a susceptibility locus for schizophrenia on chromosome 6pter-p22. Nat. Genet. *10*, 41-46. PMID: 7647789

Warkentin,S., Nilsson,A., Risberg,J., Karlson,S., Flekkoy,K., Franzen,G., Gustafson,L., and Rodriguez,G. (1990). Regional cerebral blood flow in schizophrenia: repeated studies during a psychotic episode. Psychiatry Res. *35*, 27-38. PMID: 1973302

Waters,J.J., Campbell,P.L., Crocker,A.J., and Campbell,C.M. (2001). Phenotypic effects of balanced X-autosome translocations in females: a retrospective survey of 104 cases reported from UK laboratories. Hum. Genet. *108*, 318-327. PMID: 11379878

Watson,J.D., Hopkins,N.H., Roberts,J.W., Steitz,J.A., Weiner,A.M. (1987). Molecular biology of the gene. Benjamin/Cummings. Menlo Park, California.

Wax,S.D., Rosenfield,C.L., and Taubman,M.B. (1994). Identification of a novel growth factor-responsive gene in vascular smooth muscle cells. J. Biol. Chem. *269*, 13041-13047. PMID: 8175725

Wax,S.D., Tsao,L., Lieb,M.E., Fallon,J.T., and Taubman,M.B. (1996). SM-20 is a novel 40-kd protein whose expression in the arterial wall is restricted to smooth muscle. Lab Invest *74*, 797-808. PMID: 8606489

Wei,J. and Hemmings,G.P. (2000). The NOTCH4 locus is associated with susceptibility to schizophrenia. Nat. Genet. *25*, 376-377. PMID: 10932176

Westerfield,M. (1995). The zebrafish book. Guide for the laboratory use of zebrafish (Danio rerio). 3rd ed., Univ. of Oregon Press, Eugene.

Wevrick,R., Kerns,J.A., and Francke,U. (1994). Identification of a novel paternally expressed gene in the Prader-Willi syndrome region. Hum. Mol. Genet. *3*, 1877-1882. PMID: 7849716

Williams,J., McGuffin,P., Nothen,M., and Owen,M.J. (1997). Meta-analysis of association between the 5-HT2a receptor T102C polymorphism and schizophrenia. EMASS Collaborative Group. European Multicentre Association Study of Schizophrenia. Lancet *349*, 1221. PMID: 9130948

Williams,J., Spurlock,G., Holmans,P., Mant,R., Murphy,K., Jones,L., Cardno,A., Asherson,P., Blackwood,D., Muir,W., Meszaros,K., Aschauer,H., Mallet,J., Laurent,C., Pekkarinen,P., Seppala,J., Stefanis,C.N., Papadimitriou,G.N., Macciardi,F., Verga,M., Pato,C., Azevedo,H., Crocq,M.A., Gurling,H., Owen,M.J., and . (1998). A meta-analysis and transmission disequilibrium study of association between the dopamine D3 receptor gene and schizophrenia. Mol. Psychiatry *3*, 141-149. PMID: 9577838

Wixon,J. (2000). Featured organism: Danio rerio, the zebrafish. Yeast *17*, 225-231. PMID: 11025533

Wolf,E., Kim,P.S., and Berger,B. (1997). MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci. *6*, 1179-1189. PMID: 9194178

World Health Oragnization. (1992). The ICD-10 classification of mental health and behavioural disorders. World Health Organization.

Wong,P., Myal,Y., Shui,R., and Tenniswood,M. (1993). Identification and cloning of a new category of DNA fragments which are poorly represented in human genomic libraries. Biochem. Biophys. Res. Commun. *190*, 453-461. PMID: 8381277

Wootton,J.C. and Federhen,S. (1996). Analysis of compositionally biased regions in sequence databases. Methods Enzymol. *266*, 554-571. PMID: 8743706

Wright,F.A., Lemon,W.J., Zhao,W.D., Sears,R., Zhuo,D., Wang,J.P., Yang,H.Y., Baer,T., Stredney,D., Spitzner,J., Stutz,A., Krahe,R., and Yuan,B. (2001). A draft annotation and overview of the human genome. Genome Biol. *2*, RESEARCH0025. PMID: 11516338

Wright,I.C., Ellison,Z.R., Sharma,T., Friston,K.J., Murray,R.M., and McGuire,P.K. (1999). Mapping of grey matter changes in schizophrenia. Schizophr. Res. *35*, 1-14. PMID: 9988836

Wu,Q. and Krainer,A.R. (1996). U1-mediated exon definition interactions between AT-AC and GT-AG introns. Science *274*, 1005-1008. PMID: 8875927

Wu,X.Q., Gu,W., Meng,X., and Hecht,N.B. (1997). The RNA-binding protein, TB-RBP, is the mouse homologue of translin, a recombination protein associated with chromosomal translocations. Proc. Natl. Acad. Sci. U. S. A *94*, 5640-5645. PMID: 9159125

Wu,X.Q., Lefrancois,S., Morales,C.R., and Hecht,N.B. (1999). Protein-protein interactions between the testis brain RNA-binding protein and the transitional endoplasmic reticulum ATPase, a cytoskeletal gamma actin and Trax in male germ cells and the brain. Biochemistry *38*, 11261-11270. PMID: 10471275

Wu,X.Q., Petrusz,P., and Hecht,N.B. (1999). Testis-brain RNA-binding protein (Translin) is primarily expressed in neurons of the mouse brain. Brain Res. *819*, 174-178. PMID: 10082876

Wu,X.Q. and Hecht,N.B. (2000). Mouse testis brain ribonucleic acid-binding protein/translin colocalizes with microtubules and is immunoprecipitated with messenger ribonucleic acids encoding myelin basic protein, alpha calmodulin kinase II, and protamines 1 and 2. Biol. Reprod. *62*, 720-725. PMID: 10684815

Wyatt,R.J., Alexander,R.C., Egan,M.F., and Kirch,D.G. (1988). Schizophrenia, just the facts. What do we know, how well do we know it? Schizophr. Res. *1*, 3-18. PMID: 3154503

Xu,Y., Einstein,J.R., Mural,R.J., Shah,M., and Uberbacher,E.C. (1994). An improved system for exon recognition and gene modeling in human DNA sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. *2*, 376-384. PMID: 7584416

Yin,X.Y., Grove,L.E., and Prochownik,E.V. (2001). Mmip-2/Rnf-17 enhances c-Myc function and regulates some target genes in common with glucocorticoid hormones. Oncogene *20*, 2908-2917. PMID: 11420703

Yu,Y. and Bradley,A. (2001). Mouse genomic technologiesengineering chromosomal rearrangements in mice. Nat. Rev. Genet. *2*, 780-790. PMID: 11584294

Zhang,Y., Mattjus,P., Schmid,P.C., Dong,Z., Zhong,S., Ma,W.Y., Brown,R.E., Bode,A.M., Schmid,H.H., and Dong,Z. (2001). Involvement of the acid sphingomyelinase pathway in uva-induced apoptosis. J. Biol. Chem. *276*, 11775-11782. PMID: 11278294

Zhang,Z., Ren,J., Stammers,D.K., Baldwin,J.E., Harlos,K., and Schofield,C.J. (2000). Structural origins of the selectivity of the trifunctional oxygenase clavaminic acid synthase. Nat. Struct. Biol. *7*, 127-133. PMID: 10655615

Appendix I: Sequences

Accession number[a]	Organism[b]	Description[c]	Type[d]	Length[e]
AJ310543	Human	EGLN1	N	4045
CAC42509	Human	EGLN1	P	426
AJ310544	Human	EGLN2	N	2110
CAC42510	Human	EGLN2	P	407
AJ310545	Human	EGLN3	N	2740
CAC42511	Human	EGLN3	P	239
AJ310546	Mouse	EGLN1	N	2441
CAC42515	Mouse	EGLN1	P	367
AJ310547	Mouse	EGLN2	N	2097
CAC42516	Mouse	EGLN2	P	419
AJ310548	Mouse	EGLN3	N	2656
CAC42517	Mouse	EGLN3	P	239
AJ271361	*Fugu*	WNT2, FRANK1, CFTR, FRANK2 genomic sequence.	N	60726
Q9I8E1	*Fugu*	FRANK2	P	1596
Q9I8E2	*Fugu*	CFTR	P	1511
Q9I8E3	*Fugu*	FRANK1	P	476
Q9I8E4	*Fugu*	WNT2	P	28
AF187040	Mouse	Translin associated factor X – TRAX	N	2410
Q9QZE7	Mouse	Translin associated factor X – TRAX	P	290
#	Human	DISC1 S1 splice form	N	2206
#	Human	DISC1 S1 splice form	P	662
#	Human	DISC1 S2 splice form	N	1986
#	Human	DISC1 S2 splice form	P	678
#	Human	DISC1 E3 splice form	N	1125
#	Human	DISC1 E3 splice form	P	375
#	*Fugu*	DISC1 L1 splice form	N	1607
#	*Fugu*	DISC1 L1 splice form	P	808
#	*Fugu*	DISC1 S1 splice form	N	1305
#	*Fugu*	DISC1 S1 splice form	P	435
#	*Fugu*	DISC1 S2 splice from	N	1260
#	*Fugu*	DISC1 S2 splice form	P	420
#	Zebrafish	DISC1 (partial)	N	1784
#	Zebrafish	DISC1 (partial)	P	595
#	Mouse	DISC1	N	2556
#	Mouse	DISC1	P	852
#	*Fugu*	EGLN1, TRAX, DISC1 genomic sequence.	N	44905
#	*Fugu*	EGLN1 (partial)	N	803
#	*Fugu*	ELGN1 (partial)	P	283
#	*Fugu*	TRAX	N	1185
#	*Fugu*	TRAX	P	280

Table I.i; Sequences submitted to public sequence databases. All submissions were to the EMBL database (section 2.11.1). **(a)** Accession number of entry. '#' indicates that the sequence has not yet been submitted. Sequences that have not yet been submitted to the appropriate databases will be submitted prior to the submission of relevant manuscripts to journals. **(b)** Organism the sequence was derived from. **(c)** Brief description of the sequence.

(d) 'N' indicates the sequence is nucleic acid, 'P' indicates protein. **(e)** The length of the sequence.

Assembly start[a]	Assembly end[b]	Sequence[c]	Sequence start[d]	Sequence end[e]	Orientation[f]
1	138056	AL117352_0	1	138056	
138057	259538	AL445524_1	63911	185392	
1	4628	CG_13	1	4628	
4629	4763	AC011655_15	8695	8561	C
4764	15585	CG_13	4757	15578	
15586	19119	AC011655_18	25224	21691	C
19120	26645	CG_13	19112	26637	
26646	29155	AL359543_1	1656	4165	
29156	29276	AC011655_18	11667	11547	C
29277	33146	AL359543_2	71	3940	
33147	34134	AC011655_18	7674	6687	C
34135	38368	AL359543_2	4930	9163	
38369	38590	AC011655_18	2452	2231	C
38591	38618	AL359543_3	146	173	
38619	40304	CG_13	38614	40299	
40305	43383	AL359543_3	1860	4938	
43384	44204	AC011655_13	7269	6449	C
44205	44208	CG_13	44204	44207	
44209	50648	AC011655_13	6444	5	C
50649	52544	AL359543_6	1931	36	C
52545	54624	AC011655_14	1772	3851	
54625	54946	AL359543_5	11396	11075	C
54947	55915	CG_13	54949	55917	
55916	60523	AL359543_5	10105	5498	C
60524	60529	CG_13	60530	60535	
60530	60533	AC011655_5	368	371	
60534	60565	AL359543_5	5490	5459	C
60566	60725	AC011655_5	404	563	
60726	60729	CG_13	60732	60735	
60730	60745	AC011655_5	568	583	
60746	60753	CG_13	60752	60759	
60754	61385	AC011655_5	592	1223	
61386	61391	AL359543_5	4639	4634	C
61392	62819	CG_13	61398	62825	
62820	63783	AL359543_5	3206	2243	C
63784	63805	CG_13	63790	63811	
63806	66017	AL359543_5	2221	10	C
66018	70562	CG_14	205	4749	
70563	71296	AL359543_7	2792	3525	
71297	73021	CG_14	5482	7206	
73022	77785	AL359543_9	33514	28751	C
77786	83211	CG_15	575	6000	
83212	93551	AL359543_9	23326	12987	C
93552	93580	CG_17	1223	1251	
93581	95869	AL359543_9	12958	10670	C
95870	95984	CG_18	1321	1435	
95985	106535	AL359543_9	10555	5	C
106536	106796	CG_19	5228	5488	

106797	110508	AL359543_8	6070	2359	C
110509	112270	CG_20	31	1792	
112271	112864	AL359543_8	598	5	C
112865	115021	CG_20	2386	4542	
115022	115389	AL359543_4	14276	13909	C
115390	115402	CG_21	125	137	
115403	116898	AL359543_4	13896	12401	C
116899	116938	CG_21	1633	1672	
116939	118289	AL359543_4	12362	11012	C
118290	121069	CG_22	188	2967	
121070	122646	AL359543_4	8233	6657	C
122647	125677	CG_23	1403	4433	
125678	127452	AL450284_0	4514	6288	
127453	129300	AL359543_4	1852	5	C
129301	177669	AL450284_0	8136	56504	
177670	177706	CG_31	88	124	
177707	177806	AL450284_0	56541	56640	
177807	181743	CG_31	224	4160	
181744	187007	AL450284_0	60574	65837	
187008	187024	CG_32	1654	1670	
187025	187375	AL450284_0	65854	66204	
187376	303797	AL136171_0	133872	17451	C
303798	304074	CG_35	192	468	
304075	318086	AL136171_0	17176	3165	C
318087	319817	CG_36	10188	11918	
319818	321250	AL136171_0	1437	5	C
321251	323369	CG_36	13351	15469	
323370	325634	AL445283_1	4	2268	
325635	325848	AL161743_0	97589	97376	C
325849	330610	AL445283_2	6	4767	
330611	334825	AL445200_16	16595	12381	C
334826	337851	AL445283_3	3061	6086	
337852	339975	AL445200_16	9358	7235	C
339976	340041	AL445283_3	8211	8276	
340042	340229	AL445200_16	7169	6982	C
340230	341057	AL445283_4	27	854	
341058	341483	AL445200_16	6153	5728	C
341484	341492	AL445283_4	1280	1288	
341493	341495	CG_36	33592	33594	
341496	343024	AL445200_16	5714	4186	C
343025	344300	AL445283_4	2821	4096	
344301	344380	AL445200_16	2909	2830	C
344381	344381	AL161743_0	78843	78843	C
344382	345057	AL445283_4	4178	4853	
345058	346879	AL445200_16	2153	332	C
346880	347110	AL445283_5	893	1123	
347111	347121	AL445200_16	100	90	C
347122	347503	AL445283_5	1135	1516	
347504	347811	AL445200_13	2913	2606	C
347812	348237	AL445283_5	1826	2251	
348238	350405	AL445200_13	2179	12	C
350406	350594	AL445283_6	603	791	
350595	358750	AL445200_12	8252	97	C
358751	359308	AL445283_20	569	12	C
359309	361554	AL445200_11	2458	213	C
361555	362150	AL445283_19	614	19	C
362151	363377	CG_36	54252	55478	
363378	365723	AL445200_10	7699	5354	C

365724	366120	AL445283_16	588	192	C
366121	366133	CG_36	58223	58235	
366134	371054	AL445200_10	4939	19	C
371055	371634	AL445283_15	928	349	C
371635	377625	AL445200_9	8242	2252	C
377626	377807	AL445283_21	989	808	C
377808	377811	CG_36	69907	69910	
377812	379853	AL445200_9	2065	24	C
379854	382711	AL161743_0	43376	40519	C
382712	383077	AL445283_33	4612	4247	C
383078	387917	CG_37	2862	7701	
387918	391862	AL445200_8	7320	3376	C
391863	392779	CG_37	11647	12563	
392780	393971	AL445200_8	2458	1267	C
393972	394739	CG_37	13756	14523	
394740	395178	AL445200_8	498	60	C
395179	397106	CG_37	14963	16890	
397107	397480	AL445200_14	266	639	
397481	402473	AL161743_0	25755	20763	C
402474	403230	AL445283_28	2932	2176	C
403231	414046	AL445200_15	2762	13577	
414047	414101	AL445283_26	109	55	C
414102	414372	CG_38	12634	12904	
414373	415169	AL445200_15	13904	14700	
415170	420353	AL161743_0	8065	2882	C
420354	420357	CG_38	18878	18881	
420358	420605	AL161743_0	2877	2630	C
420606	420618	CG_38	19130	19142	
420619	422138	AL445200_7	3079	1560	C
422139	424075	CG_38	20663	22599	
424076	424082	AL445283_23	3760	3754	C
424083	424898	AL445200_6	3749	2934	C
424899	424929	AL445283_23	2937	2907	C
424930	424937	CG_38	23451	23458	
424938	427037	AL445283_23	2898	799	C
427038	427093	AL445200_6	795	740	C
427094	427232	CG_38	25615	25753	
427233	427768	AL445200_6	600	65	C
427769	427780	CG_38	26290	26301	
427781	427812	AL445200_6	52	21	C
427813	430252	CG_38	26334	28773	
430253	433484	AL445283_24	3255	24	C
433485	455441	AL35305_1	1217	21849	

Table I.ii; A hybrid assembly of human genomic sequence over the *EGLN1*, *TRAX* and *DISC1* genomic region. The coordinates and sequence versions can be used to recreate the assembly of contiguous human genomic sequence. Assembly coordinates refer to nucleotide positions within the contiguous assembled sequence. Sequence coordinates refer to the nucleotide positions within fragments of individual clone sequences. **(a)** Start point of a sequence fragments contribution to the assembly in assembly coordinates. **(b)** End point of a sequence fragments contribution to the assembly in assembly coordinates. **(c)** The sequence identifier. The letters and numbers to the left of the underscore indicate the EMBL accession number of a clone sequence (beginning AC or AL) or indicate a contribution by a Celera

Genomics (CG) sequence scaffold GA_x8YCEG9. Numbers after the underscore indicate which fragment of sequence is referred to (section 4.#). **(d)** Start point of a sequence fragments contribution to the assembly in sequence fragment coordinates. **(e)** End point of a sequence fragments contribution to the assembly in sequence fragment coordinates. **(f)** Orientation of the sequence fragment in the assembled sequence contig, 'C' indicates a complementary orientation (inverted).

Reaction set[a]	Forward	Reverse
1	CCCTGTGGCTACTGAGGTC	CACCCAGGGGGTAAGTGAC
2	TCCTTTTCCCACTTTTATCCA	CCAACATTATCAAAGCCCAGA
3	AGGGTGGGTACATGTGTCTCTT	ACTCCTACCTGGGGGCATT
4	GAAGAGCCTGTTTGCCATGT	CAGAAATCTGTGCCCAAGAA
5	GCAGATTTCTTCCTCCATTGC	TGGCAGGTAAGAGGGATTCA
6	TCCTGTCACAGAGCCTTCAG	TCTCCTAGTCACCCTTCGAAAA
7	AGTTTTCAAGCACCCTCCCT	TCTTAAAAGCAAACAGGGCTTC
8	AAATGCTGCTGAAATGGCAATCTTT	TGCAAAACCTCACAAAATCAA
9	GGATTTTCAGGAATGAAAGTTGA	GCCTCAATGGAAAGTGGATTAG
10	TGCTTTGAGGAGCTAGTGTTTG	ATTTGGGGAGACTGGAAATGTA
11	TTTTTGGTGTATTAACCCATTATGA	GGGAGAATCACTTCACTTCCAA
12	TCACCCAACTCCACTGTTTTC	GTGCCATGCTCTCACTATTGTC
13	TTGTACTGAGAAAGAGGATGCAAT	TTGGGTCATTAATCCCAATCTT
14	TGCAGATTGATTTGCAGGTT	TCTCCCTTCGCTGACTCTCT
15	CTACTTTTTTTCATTGTGTGTTTCAC	AACTATCATATCCCCTGTGACCT
A	TGATTTCGGATTCAGGAAGG	TCCTGATTAGGCATGGAAGG
B	TGAAGGAGGCACCTAGTGGT	TGCACCCTAATGCCTTAAGAA
C	CCCTAGTGTGGCTCTTCTGG	AAAGTCAGAGGCTGGAGCAAA

Table I.iii; Oligonucleotides used to map the 5' end of *DISC2* by RT-PCR. Forward and reverse indicate the relative orientation of oligonucleotides aligned with the human genomic sequence contig (section 4.3). Nucleotide sequences are shown 5' (left) to 3' (right) **(a)** The position of each reaction set in genomic sequence is indicated in section 7.4.

Probe name	Primer reference	Primer sequences
D1P	M935	CGGGATTGCTTACCACCTG
	M682	CCATCTGCCTCCGGTTTC
D79P	A1504	TCAGAGCAGTTTGCCATGAGC
	A1505	ACCAGTGGGTCAGGTCGCAG
D80P	S330	CTGAGAAGGGAAATAGAGGAG
	S328	GAGCGACCAAAATCAGTTAG
D95P	S323	TGGCTGTTCCACTGCCTTCTG
	S324	GCTGTCATTATTTTAGTTGGATAG

Table I.iv; Human derived *DISC1* probes for cross-species hybridisation. The oligonucleotides indicated were used to PCR amplify from human cDNA pools. The PCR products were purified (section 2.4.4) and labelled (section 2.8.2) prior to DNA:DNA hybridisation (section 2.10.2).

Probe name	Primer reference	Primer sequences
T-P1	U301	GGCCTTTAAATCATTTCAGCAG
	U302	TCTTCCATATCAGGAGCACTTGT
T-P2	U303	TGGAGACTGAGAGTCACACCTG
	U304	TCACTTCAAAGGGGGTATCAAT
T-P3	V103	TGTAGTAATGGCTCGATGGAACTG
	V102	CCCACATAACCAAAGAAGAGAAGG
T-P4	V104	TTTTCAACACTTCATCAAAACACG
	V105	CATAACAAGCATTCTCCACTTTGG

Table I.v; Human derived *TRAX* probes for cross-species hybridisation. The oligonucleotides indicated were used to PCR amplify from human cDNA pools. The PCR products were purified (section 2.4.4) and labelled (section 2.8.2) prior to DNA:DNA hybridisation (section 2.10.2).

Appendix II:

Software developed during the course of this thesis

II.i The sgrab Perl module

```
package sgrab;
#
# Martin Taylor 2001.
#
use strict;
use Exporter;
our (@ISA,@EXPORT,$VERSION);
$VERSION=1;
@ISA=qw(Exporter);
@EXPORT=qw(new spew grab spcw2filc getlength);
sub new {
        my ($subname,$fi,$beg,$end)=@_;
        my $seqObj = {
                "file_name" => "$fi",
                "beg" => "$beg",
                "end" => "$end",
                "name" => "null",
                "rawseq" => "null",
                "linelength" => "null",
                "lines" => "null",
        };
        bless $seqObj, 'sgrab';
        return $seqObj;
}

sub grab {
        my $subseq = shift;
        my $fi = $subseq->{"file_name"};
        my $beg = $subseq->{"beg"};
        my $end = $subseq->{"end"};
        open SEQFILE, "<$fi" || die "file $ARGV[0] not readable\n";
        my ($fist,@format,$template,$outlines,$offset,@splitfirst);
        my $firstline = <SEQFILE>;
        my $secondline = <SEQFILE>;
        my $headerLen = length $firstline; #get header length
        my $defaultLen = length $secondline; # define the standard line length
        @splitfirst = split /\s+|:|\.|;|#/, $firstline;
        $splitfirst[0] =~ s/^>//;
        $defaultLen -= 1; #don't want to count CR at the moment.
        my $seqoffset = $end - $beg;
        my $crMod = int ($beg / ($defaultLen)); # how many CRs to the start point
        my $ADDRESS = $crMod + $beg + $headerLen -1;
        # if the length of 1st line and last line is >= line length, there will
        # be an additional CR to count.
        my $relpos = ($ADDRESS % $defaultLen);
        my $relposE = ($end % $defaultLen);
        $outlines = int ($seqoffset / $defaultLen);
        if ($seqoffset % $defaultLen==0){$outlines--;}
        $offset = $seqoffset + $outlines ;
        if (($relpos + $relposE) > $defaultLen){ $offset += 1;} # that troublesome CR!
        $outlines += 2;
        seek SEQFILE, $ADDRESS,0; # move file pointer to start point
        read SEQFILE, $fist, $offset; # grab the sequence chunk
        $fist =~ s/\n//g; # This seems prety fast even for 10s of Mb!
        $subseq->{"rawseq"}=$fist;
        $subseq->{"linelength"}=$defaultLen;
        $subseq->{"lines"}=$outlines;
        $subseq->{'name'}=$splitfirst[0];

}
sub spew {
```

```
        my $subseq = shift;
        my $template = "a$subseq->{'linelength'}"x$subseq->{'lines'};
        my @format = unpack $template, $subseq->{'rawseq'};
        print ">",$subseq->{'name'},"\n";
        print (join "\n", @format);
        if ((length (pop @format)) % $subseq->{'linelength'} !=0){print "\n";}
}
sub spew2file {
        my $subseq = shift;
        my $fileout = shift;
        my $template = "a$subseq->{'linelength'}"x$subseq->{'lines'};
        my @format = unpack $template, $subseq->{'rawseq'};
        open TOSPEW, ">$fileout" || die "Cant write file $fileout\n";
        print TOSPEW ">",$subseq->{'name'},"\n";
        print TOSPEW (join "\n", @format);
        if ((length (pop @format)) % $subseq->{'linelength'} !=0){print TOSPEW "\n";}
        close TOSPEW;
}
sub getlength {
        my $seqfile = shift;
        my ($headlen,$secondline,$bytesize,$minustail,$taillen,$lines);
        my ($secondchoped);
        $bytesize = -s $seqfile;
        open SQFILE, "<$seqfile" || die "cant open file $seqfile\n";
        $headlen = length <SQFILE>;
        $secondline = <SQFILE>;
        $secondchoped  = $secondline;
        $secondchoped =~ s/\W+$//;
        $minustail = length $secondchoped;
        $taillen = (length $secondline) - $minustail;
        $lines = int (($bytesize - $headlen)/($taillen + $minustail));
        if ((($bytesize - $headlen) % $minustail) != 0){$lines++;}
        my $seqlen = ($bytesize - $headlen) - ($taillen * $lines);
        return $seqlen;
}
```

II.ii Using the sgrab Perl module

```
#!/bin/perl -w
use strict;
use sgrab;
my $ob = sgrab->new($ARGV[0],$ARGV[1],$ARGV[2],$ARGV[3]);
sgrab::grab($ob);
sgrab::spew($ob);
```

II.iii Annotation anchored global sequence alignment

```
#!/bin/perl -w
use strict;
use sgrab;
my ($lim,@ordinate,@coords,$logt);
open ANCHORS, "<$ARGV[0]";
foreach $lim (<ANCHORS>){
        if ($lim =~ /^#/){next;}
        if ($lim =~ s/^>//){
                chomp $lim;
                push @ordinate, $lim;
        }
        elsif ($lim =~ /\d+/){
                chomp $lim;
                @tsp = split /\s+/, $lim;
                push @qu, $tsp[0];
                push @sbj, $tsp[1];
                $noanchors++;
        }
}

unshift @qu, "1";
unshift @sbj, "1";
```

```
push @qu, sgrab::getlength($ordinate[0]);
push @sbj, sgrab::getlength($ordinate[1]);

if ($#qu != $#sbj){die "Bad error, anchors not balanced!\n";}
my $lop=0;
while ($lop <= $noanchors){
        $qout = $ordinate[0].".aagsalTMP.".$lop;
        $sout = $ordinate[1].".aagsalTMP.".$lop;
        push @qoutlist, $qout;
        push @soutlist, $sout;
        $qbeg = $qu[$lop]-1;
        $qend = $qu[($lop+1)]-2;
        my $qob = sgrab->new($ordinate[0],$qbeg,$qend);
        sgrab::grab($qob);
        sgrab::spew2file($qob,$qout);
        $sbeg = $sbj[$lop]-1;
        $send = $sbj[($lop+1)]-2;
        my $sob = sgrab->new($ordinate[1],$sbeg,$send);
        sgrab::grab($sob);
        sgrab::spew2file($sob,$sout);
        $lop++;
}

my $co=0;
foreach $lim (@qoutlist){
        print "$lim $soutlist[$co]\n";
        $logt = eval {
                `avid -nm=both $lim $soutlist[$co]`
        };
        push @TMPs, "$lim\_$soutlist[$co].out";
        $co++;
}
&fuseAnchorAln(@TMPs);

sub fuseAnchorAln{
        my @tp = @_;
        $outfile = "aagsal.outfile";
        my (@splt,$snd,$stella);
        my ($colo,$colt)=(0,0);
        open BAR, ">$outfile";
        foreach $fileTMP (@tp){
                print "$fileTMP\n";
                open CURRTMP, "<$fileTMP" || die "failed to open $fileTMP\n";
                while (<CURRTMP>){
                @splt=split /\s+/, $_;
                if ($splt[0] =~ /\d+/){
                        $colo++;
                        $splt[0] = $colo;
                }
                        elsif($splt[1] eq "|"){
                                $splt[1] = "    |";
                        }
                $snd=pop @splt;
                if ($snd =~ /\d+/){
                $colt++;
                push @splt, $colt;
                }
                elsif($snd =~/\|/){push @splt, "  |";}
                $stella = join " ", @splt;
                print BAR "$stella\n";
                }
                close CURRTMP;
        }
}
```

II.iv VistaWrap – an example wrapper script

```
#!/bin/perl -w
use Getopt::Std;
use Usage;usage();
use strict;
our %opts;
our ($alignfile,$annotate);
my ($key,%assoc);
# Stuff for java to find vista
$ENV{CLASSPATH}="/packages/vista/Vista.jar:/packages/vista/retepPDF2.jar";
getopts('t:o:a:r:b:m:n:p:c:', \%opts);
%assoc= (
        t => 'title',
        o => 'output',
        a => 'alignfile',
        f => 'format',
        r => 'id',
        b => 'blocksize',
        m => 'minIDplot',
        n => 'annotate',
        c -> 'coordin'
);
if(!defined $opts{a}){print "You must define an alingment file\n";die;}
my $tmpfile = "vistawrap.tempplot";
if (defined $opts{p}){$tmpfile = $opts{p};}
open FOUT, ">$tmpfile" || die "Cant open temp file\n";
our $title="null";
our $output="vistaout.pdf";
our $format="GLASS";
our $species1="human";
our $species2="mouse";
our $id=70;
our $blocksize=100;
our $minIDplot=50;
our $coordin=1;

foreach $key (keys %opts){
        no strict 'refs';
        ${$assoc{$key}} = $opts{$key};
}
$coordin--;
my @spec=($species1,$species2);
if (defined $opts{n}){$annotate=$opts{n};}
print FOUT "TITLE $title\n";
print FOUT "OUTPUT $output\n";
print FOUT "ALIGN $alignfile $format\n";
print FOUT "\tSEQUENCES $species1 $species2\n";
print FOUT "\tREGIONS $id $blocksize\n";
print FOUT "\tMIN $minIDplot\n";
print FOUT "END\n";
print FOUT "COORDINATE $spec[$coordin]\n";
if (defined $annotate){print FOUT "GENES $annotate\n";}
print FOUT "PAPER A4\n";
close FOUT;
eval {`java Vista $tmpfile`;};

print "Your PDF file $output is ready\n";


__END__

=head1 NAME

vistawrap

=head1 SYNOPSIS

vistawrap -a I<alignmentfile> [options]

=head1 DESCRIPTION
```

```
A complete wrapper for the Vista program, generating the configuration file on the
fly. The only option that must
be given is -a to define a GLASS format alignment (also produced by AVID). A PDF
format graphic showing patterns o
f conservation in a pairwise alignment is the eventual output of this program.

Vista can be found at http://www-gsd.lbl.gov/vista/ and any use should reference the
original paper: Mayor et al.,
 Bioinformatics 16:1046 (2000).


=head1 OPTIONS

B<-a> I<file>           Defines alingment file. This is mandatory.
                        Expects GLASS format generated by AVID.

B<-t> I<title>          Give a title for the image. (Default= null).

B<-o> I<output>         Name the output file.
                        (Default= vistaout.pdf).

B<-r> I<int>            Minimum ID for coloring alignment.
                        (Dcfault 70%).

B<-b> I<int>            Size of sliding window for average
                        calculation. (Default 100).

B<-m> I<int>            Minimum ID to plot. (Default 50).

B<-n> I<file>           Name of annotation file in "simple" format.
                        If excluded no annotation is used.

B<-p> I<file>           Give alternate name for temporary Plotfile.

B<-c> I<int>            Which sequence shoud be the base (coordinate)
                        sequence. (Default 1) 2 is the other valid
                        option.


=head1 By

B<Martin Taylor> 2001
martin.taylor@ed.ac.uk
=cut
```

II.v grab

```
#!/bin/perl
if ($ARGV[0] =~ /^\-/){
        $slow++;
        if ($ARGV[0] !~ /s/){ $splice++;}
        if ($ARGV[0] =~ /m/){ $mask++;}
        $beg = $ARGV[1]; $end = $ARGV[2];
}
else{
        $beg = $ARGV[0];
        $end = $ARGV[1];
}

# Get the sequence, parse and shove into @
while (<STDIN>){
        if (/^>/){ $in++; $ti = $_; chomp $ti; next; }
        if ($in >= 1) {
                @seqin[$cnt] = $_;
                $cnt++;
        }
}
$fulseq = join ("", @seqin);
$fulseq =~ s/\W+//g;
@stream = split //, $fulseq;
$len = @stream;
if ($slow == 1){ & sandm;}
```

```
else {& sgrab;}
# Basic grab subroutine
sub sgrab (){
        if ($end < 1){ $end = $len;}
        print "$ti#$beg\_$end\n";
        undef $con; undef $numb;
        $con = $beg - 2; $outs++;
        while (){
                $numb++;
                $con++;
                if ($con >= $end){ last;}
                $ot = uc @stream[$con];
                print "$ot";
                if ($numb % 50 == 0){ print "\n";}
        }
        if ($numb % 49 != 0){ print "\n";}
}
# Mask and splice subroutine (slower)
sub sandm (){
        print "$ti\n";
        undef $con; undef $numb;
        while (){
                if (@stream[$con] !~ /\w/){last;}
                $ot = uc @stream[$con];
                if ($con == $beg -1){ $mck++;}
                if ($con == $end){ $mck--;}
                $con++;
                if ($mask == 1){
                        if ($mck == $splice){
                                $ot = "N";
                                print "$ot";
                                $numb++;
                        }
                        else{ print "$ot"; $numb++;}
                }
                elsif ($mask == 0){
                        if ($mck == $splice){
                                print "$ot"; $numb++;
                        }
                        else {next;}
                }
                #else { next;}
                if ($numb % 50 == 0){ print "\n";}
        }
        if ($numb % 49 != 0){ print "\n";}
}
```

II.vi revcomp

```
#!/bin/perl
#
# Martin Taylor, MRC Human Genetics Unit, Edinburgh
# 2000
while (<STDIN>){
        if (/^>/){ $in++; $ti = $_; chomp $ti; next; }
        if ($in >= 1) {
                @seqin[$cnt] = $_;
                $cnt++;
        }
}
chomp(@seqin);
$fulseq = join ("", @seqin);
$fulseq =~ y/AaTtCcGgYyRrNn/TtAaGgCcRrYyNn/;
@stream = split //, $fulseq;
$len = @stream;
&rgrab;
sub rgrab (){
        print "$ti#$beg\_$end\n";
        while ($len--){
                $ot = uc @stream[$len];
                $con++;
                print "$ot";
                if ($con % 50 == 0){ print "\n";}
```

```
        }
        if ($con % 50 != 0){ print "\n";}
}
```

II.vii stfa

```
#!/bin/perl
if ($ARGV[0] eq "\-n"){ $loud++;}
foreach $fi (<STDIN>){
        if ($fi =~ /^>/){
                opendir CDIR, ".";
                @dirlst = readdir CDIR;
                closedir CDIR;
                $fc = @dirlst;
                $flist = join '##', @dirlst;
                @ti = split  (/\s+/, $fi);
                @ti[0] =~ s/^>//;
                $in++;
                $len = 0;
                $tii = @ti[0];
                #$tii = join '.', @ti[0], stfa;
                if ($flist =~ /$tii/){
                        $fcc = 0;
                        while (){
                                $ren++;
                                $tiii = join '.', $tii, $ren;
                                if ($flist !~ /$tiii/){ $tii = $tiii; $ren = 0; last;}
                        }
                }
                if ($loud == 1){
                        $arrow = " <== ";
                        print STDERR "$tii $arrow $fi\n";
                }
                open (FON, ">$tii");
                print FON ">@ti[0]\n";
                next;
        }
        if ($fi =~ /^\s+/ || $fi =~ /^#/){ next;}
        if ($in >= 1 && $fi =~ /\S+/){
                @stmp = split (/\s+/, $fi);
                $len = split (//, @stmp[0]);
        }
        if ($in >= 1 && $len >= $leno){print FON "@stmp[0]\n";}
        if ($in >= 1 && $len < $leno){
                if ($len >= 1){ print FON "@stmp[0]\n";}
                close FON;
                $in = 0;
        }
}
```

Appendix III: Manuscripts published during the course of this thesis

Research manuscripts

Taylor MS. Characterisation and comparative analysis of the *EGLN* gene family. Gene. 2001 Sep 5;275(1):125-32.

Davidson H, Taylor MS, Doherty A, Boyd AC, Porteous DJ. Genomic sequence analysis of *Fugu* rubripes *CFTR* and flanking genes in a 60 kb region conserving synteny with 800 kb of human chromosome 7. Genome Res. 2000 Aug;10(8):1194-203.

Millar JK, Wilson-Annan JC, Anderson S, Christie S, Taylor MS, Semple CA, Devon RS, Clair DM, Muir WJ, Blackwood DH, Porteous DJ. Disruption of two novel genes by a translocation co-segregating with schizophrenia. Hum Mol Genet. 2000 May 22;9(9):1415-23.

Devon RS, Taylor MS, Millar JK, Porteous DJ. Isolation and characterization of the mouse translin-associated protein X (*Trax*) gene. Mamm Genome. 2000 May;11(5):395-8.

Review manuscripts

Taylor MS. More biology from the sequence. Genome Biol. 2001; 2(8):REPORTS4018.

Semple CA, Taylor MS, Ballereau S. The meso-genomic era. Genome Biol. 2001;2(7):REPORTS4015.

Reprints of the research manuscripts are included in the subsequent pages.

Manuscripts are reproduced with the permission of copyright owners.

ELSEVIER

Gene 275 (2001) 125–132

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Characterization and comparative analysis of the EGLN gene family

Martin S. Taylor*

Medical Genetics Section, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK

Received 10 April 2001; received in revised form 4 July 2001; accepted 25 July 2001
Received by S. Salzberg

Abstract

Rat *Sm-20* is a homologue of the *Caenorhabditis elegans* gene *egl-9* and has been implicated in the regulation of growth, differentiation and apoptosis in muscle and nerve cells. Null mutants in *egl-9* result in a complete tolerance to an otherwise lethal toxin produced by *Pseudomonas aeruginosa*. This study describes the conserved Egl-Nine (EGLN) gene family of which rat SM-20 and *C. elegans* Egl-9 are members and characterizes the mouse and human homologues. Each of the human genes (*EGLN1*, *EGLN2* and *EGLN3*) are of a conserved genomic structure consisting of five coding exons. Phylogenetic analysis and domain organization show that *EGLN1* represents the ancestral form of the gene family and that *EGLN3* is the human orthologue of rat *Sm-20*. The previously observed mitochondrial targeting of rat SM-20 is unlikely to be a general feature of the protein family and may be a feature specific to rats. An EGLN gene is unexpectedly found in the genome of *P. aeruginosa*, a bacterium known to produce a toxin that acts through the Egl-9 protein. The pathogenic bacterium *Vibrio cholerae* is also shown to have an EGLN gene suggesting that it is an important pathogenicity factor. These results provide new insights into host–pathogen interactions and a basis for further functional characterization of the gene family and resolve discrepancies in annotation between gene family members.

Keywords: Smooth muscle-20; Egg laying-9; Phylogeny; Apoptosis; Mitochondria

1. Introduction

Egl-9 was originally isolated as the gene responsible for an egg laying defective phenotype of the nematode worm *Caenorhabditis elegans* (Trent et al., 1983). More recently it has been demonstrated that the Egl-9 gene product is a target or mediator of a diffusible toxin produced by some strains of *Pseudomonas aeruginosa* (Darby et al., 1999). In *C. elegans*, the toxin induces rapid neuromuscular paralysis and loss of function mutations in *egl-9* confer a strong resistance to this toxicity.

SM-20, a vertebrate homologue of Egl-9, has been implicated in the differentiation and growth regulation of muscle cells in the rat (Wax et al., 1994; Moschella et al., 1999) and has subsequently been shown to be necessary for nerve growth factor-dependent survival of neurons (Lipscomb et al., 1999). Overexpression of SM-20 is able to induce apoptotic cell death in neurons (Lipscomb et al., 1999). Work involving the activation of temperature-sensitive p53 and the overexpression of rat SM-20 in cells lacking functional p53 (Madden et al., 1996) has implicated SM-20 as a downstream mediator of p53 signalling. SM-20-induced cell death has been shown to be accompanied by caspase-3 activation and inhibition of caspase activity prevents SM-20 induction of apoptosis (Lipscomb et al., 2001).

Further investigation of this neural, apoptotic induction by Lipscomb et al. (2001) has led to a convincing demonstration of mitochondrial targeting of SM-20. The mitochondrial targeting signal of SM-20 has been mapped by fusion and deletion experiments to the first 25 amino acids of the protein (Lipscomb et al., 2001). This 25 amino acid region was observed to be rich in hydroxylated and basic amino acids and devoid of acidic amino acids, as is typical for mitochondrial targeting sequences (Hurt and Schatz, 1987). It was also shown by deletion of the mitochondrial targeting sequence that mitochondrial localization is not necessary for the role of SM-20 in apoptotic induction

Abbreviations: BAC, bacterial artificial chromosome; BLASTN, basic local alignment search tool-N; bp, base pair; cDNA, complementary DNA; Egl-9, egg laying-9; EGLN, egg laying nine; EGLN1, egg laying nine-1; EGLN2, egg laying nine-2; EGLN3, egg laying nine-3; EMBL, European Molecular Biology Laboratory; EST, expressed sequence tag; FIS, full insert sequence (of a cDNA clone); GSS, genomic survey sequence; HMM, hidden Markov model; HTG, high throughput genomic; HUGO, human genome organization; mRNA, messenger ribose nucleic acid; nt, nucleotide; PAC, P1 artificial chromosome; SM-20, smooth muscle-20; TBLASTN, T-basic local alignment search tool-N; UTR, untranslated region

* Tel.: +44-131-651-1084; fax: +44-131-651-1059.
E-mail address: martin.taylor@ed.ac.uk (M.S. Taylor).

0378-1119/01/$ - see front matter
PII: S0378-1119(01)00633-3

(Lipscomb et al., 2001). In the absence of a mitochondrial targeting signal, SM-20 was observed to localize to both the cytosol and nucleus.

Caenorhabditis elegans Egl-9 (O45918), rat *Sm-20* (AAG33965), human *C1orf12* and *SCAND2* genes have previously been described. It is, however, apparent from BLAST (Altschul et al., 1990) homology searching that there are multiple other homologous sequences represented in the EST and genomic sequence databases. This study set out to identify and characterize the human and mouse homologues and investigate the evolutionary history of this interesting and little understood gene family.

2. Methods

2.1. EST clustering

The clustering of ESTs was seeded by an initial search using *C. elegans* Egl-9 and rat SM-20 amino acid sequences as queries against mouse or human subsets of the EMBL EST dataset (http://www.ebi.ac.uk/embl/). The Blast2 (Altschul et al., 1997) implementation of TBLASTN was the search algorithm used. All EST matches with a TBLASTN bit-score of 58 or greater were searched against the EST database of the appropriate species using BLASTN to identify all overlapping EST sequences. Matches with a BLASTN bit-score of 80 or greater were assembled into contigs using phrap (Ewing et al., 1998) with default parameters. The TBLASTN bit-score cut-off of 58 was used as it represented a boundary between the best matches and all other sequences in the distribution of scores when Egl-9 was searched against human ESTs. The BLASTN cut-off used in EST assembly was empirically determined to maximize extension of contigs and identification of homologous, non-identical sequences while minimizing matches to unrelated sequences. Consensus sequences of the phrap-generated assemblies were masked for interspersed repeat elements using RepeatMasker (Phil Green, unpublished data) with the xsmall option set and using rodent or human repeat databases as appropriate for the assembly. Masked assemblies were searched again against the appropriate EST database, all new matches with bit-scores of 80 or greater were obtained and the phrap assembly was repeated to include the new matches. This process of iterative BLASTN searching, phrap assembly and repeat masking was repeated until no new EST sequences were obtained.

The EMBLminus dataset was searched along with ESTs in the cycles of contig extension. All matches from the EMBLminus dataset were manually screened to ensure the sequences were from the expected species and that they represented transcripts rather than genomic sequence. If EMBLminus sequences met these criteria, they were also included in the phrap assembly.

All assemblies were manually curated to ensure that closely related genes were not assembled into the same contig; this was essential to separate *SCAND2* from *EGLN1* transcripts which are >90% identical in global alignment and have several stretches of 100% identity for >80 nucleotides.

2.2. Homology searching in genomic sequence

C. elegans Egl-9 and rat SM-20 were used as queries for TBLASTN searching of all publicly available human genomic sequence (from HTG, GSS and EMBLminus subsets of the EMBL database). TBLASTN was used with the Blosum62 substitution matrix and default parameters. Transcript sequences were aligned to genomic sequence by est2genome (Mott, 1997) with default parameters. Amino acid sequences were aligned to genomic sequence using TBLASTN with a 'flat' matrix (20 points for a match, −5 for a mismatch), gapping behaviour and compositional filtering switched off and the number of best hits from a region limited to one. High scoring segment pairs from the TBLASTN output were used as starting points for genewise (wise2-1-20c) (Birney, unpublished data) alignment of the amino acid sequence to genomic sequence. This strategy substantially reduced the computational time required.

2.3. Distant homology searching

Hidden Markov models (HMMs) of ClustalW (version 1.74; Thompson et al., 1994) aligned amino acid sequences were generated using the HMMER tools (http://hmmer.wustl.edu/hmmer-html/), hmmbuild and hmmcalibrate with a seed of 0 and default parameters. Unless otherwise stated, all HMMs were generated with the multi-hit local search algorithm style (option -f). Searching of protein sequence databases was carried out using hmmsearch with default parameters.

To search nucleotide databases with HMMs, estwisedb (wise2-1-20c) (Birney, unpublished data) was used when searching viral, EST and non-eukaryotic genomic sequences. For protein HMM searching of human genomic sequence, each of the amino acid sequences used to generate the alignment were used as queries for TBLASTN searching of the nucleotide database. Every hit with a bit-score of greater than 38 (a low and arbitrary cut-off) was aligned with the HMM using genewise (as described, see Section 2.2).

2.4. Phylogenetic analysis

Components of the Phylip package (Felsenstein, unpublished data) were used for tree building. A manually edited ClustalW alignment was used to generate 1000 alignment datasets through the delete-half jackknife method. Dayhoff PAM matrices were calculated for each dataset and neighbour joining was used to estimate phylogenies from each distance matrix. A consensus tree was derived using the consense function from Phylip.

3. Results and discussion

3.1. Identification of human and mouse Egl-9 homologues

Egl-9 and SM-20 homologous sequences were well represented in the EST subsection of EMBL. Assemblies representing four human genes and three mouse genes were generated from the clustered ESTs. One of the human assemblies comprising 11 ESTs and four mRNA database entries represents the previously described *SCAND2* transcript (Dupuy et al., 2000), which although derived in part from a human Egl-9 homologue is likely to be translated in an altered reading frame and consequently would not produce an Egl-9-related protein product.

Nucleotide identity between pairs of human and mouse EST assemblies strongly suggests that these three paralogous human transcripts are orthologues to the three mouse transcripts. These genes are hereafter termed *EGLN1*, *EGLN2* and *EGLN3* (HUGO accepted nomenclature) with reference to their mutual homology with Egl-Nine.

Preliminary examination of Egl-9 homology in human genomic sequence suggested that Egl-9 homologues are located at more than nine discrete chromosomal locations based on sequence-associated annotation alone. However, further investigation demonstrated that several of the fragmentary BAC clone sequences annotated to be from different chromosomal locations are likely to represent common genomic loci (Table 1 and data not shown).

Based on the clustering of genomic sequence accessions (Table 1), there are five regions in the human genome with strong homology to Egl-9 and SM-20. The *SCAND2* gene on chromosome 15 has been previously reported (Dupuy et al., 2000). Sequences from the overlapping BAC clones AC018654 and AC022073 show 65% identity in conceptual translation to rat SM-20 (data not shown). However, a lack of introns, multiple in-frame stop codons, a lack of representation in the EST databases and an Alu element insertion into the coding sequence strongly imply that this is a processed pseudogene. The remaining three regions correspond to *EGLN1*, *2* and *3* (Table 2).

The sequence of human *EGLN1*, *EGLN2* and *EGLN3*, and mouse *EGLN1*, *EGLN2* and *EGLN3* transcripts has been submitted to the public sequence databases under Accession numbers AJ310543, AJ310544, AJ310545, AJ310546, AJ310547 and AJ310548, respectively.

3.1.1. EGLN1

Human *EGLN1* (*C1orf12*) has been previously reported and its genomic structure described (Dupuy et al., 2000). In Northern blot analysis, a major band at ~5 kb and a smaller, less abundant ~2.4 kb band were observed (Dupuy et al., 2000). Assembly of 60 human ESTs results in a contiguous assembly of 4029 bp. The addition of a poly A tail and additional 5′ UTR not represented in the EST assembly is likely to correspond to the observed ~5 kb transcript. Four of the ESTs demonstrate that a second polyadenylation site

Table 1
Genomic content and location of the human EGLN gene family

Gene name	Genomic sequences[a]	Chromosome annotation[b]	Accession Map annotation[c]		
			Contig[d]	Chromosome[e]	RH map[f]
EGLN1	AL445524	1	NA	–	–
	AL358784	1	ctg13079	1	736.71
	AL117352	1q42	ctg13079	1	736.73
	AC011945	NA	ctg13079	1	736.84
	AC012242	NA	ctg13079	–	–
EGLN2	AC008537	19	ctg15547	19	235.50
	AC025769	5	ctg14739	19	236.31
	AC019337	NA	ctg17685	19	235.44
EGLN3	AC022969	14	ctg53	14	65.90
	AL358340	14	ctg53	14	65.90
	AC023450	1	ctg53	–	–
	AC084333	4	NA	–	–
SCAND2	AC048382	18	ctg12246	15	300.12
	AC016771	15	ctg12246	15	300.43
	AC087732	15	NA	–	–
ψEGLN3	AC018654	12	ctg14210	12	69.39
	AC022073	12	ctg14210	12	66.75

[a] Sequence Accession numbers for each genomic clone with significant (Section 2.2) homology to *C. elegans* Egl-9 and rat SM-20 clustered by sequence identity.
[b] Map annotation for each of the sequence accessions.
[c] Data derived from the Accession Maps project (http://genome.wustl.edu/gsc/human/Mapping/).
[d] Accession Map contig of the sequence, based on the October 7th 2000 data freeze.
[e] Chromosomal location as determined by STS content of the clone sequence.
[f] Radiation hybrid (RH) map location of the clone.

Table 2
Minimal tiling path of EST assemblies

Gene[a]	mRNA[b]	EST tiling path[c]
Human EGLN1	AF229245	AI343586, AF229245, BE669740, BE960611, AA508346, BG117305, AA370652, AA203627, AW377161, AA218859
	AF334711	
	AF277176	
	AJ227859	
	AF277174	
Human EGLN2	AK026863	BF061631, BE561402, AW957364, AA312497, AI110596, AL40033
	AK025396	
Human EGLN3	AK025273	BF726360, BF724589, R00332, AW079532, A804037
	AX035283	
	AK026918	
Mouse EGLN1		BF465779, AW610745, AA434738, BF022323, BF161755, BF782795
Mouse EGLN2		BF321818, BF301163, AW320216, AA939949, AA014327, BF730642
Mouse EGLN3		BF540183, BF142691, W14123, AI020522

[a] Mouse and human Egl-9 related genes identified by EST clustering.
[b] cDNA sequences representing the transcription of these genes.
[c] Tiling path of ESTs from each assembly, from which the complete assemblies can be derived.

is also used. This produces a transcript of 1879 bp, which is likely to represent the less abundant ~2.4 kb transcript.

Aligning the *EGLN1* EST assembly with the transcript reported by Dupuy et al. (2000) results in a predicted transcript of 7089 bp, not including polyadenylation. This is at least 2 kb longer than the major transcript detected by Northern blot. The compelling evidence from ESTs is for a major transcript that contains all of the assembled 4029 bp of sequence. The transcription initiation site proposed by Dupuy et al. (2000) is approximately 2 kb upstream of the major transcriptional start site inferred here.

Mouse *EGLN1* was assembled from a cluster of >97 ESTs into a 2835 bp contig that is incomplete at the 5′ end and demonstrates the use of at least two polyadenylation signals. Both human and mouse *EGLN1* transcripts make use of a non-consensus ATTAAA polyadenylation signal. Within the 3′ UTR of the mouse and human transcripts there is a block of approximately 300 nt showing 86% identity between species (data not shown). This is substantially greater conservation than the remainder of the 3′ UTR and may reflect a site of regulatory importance.

3.1.2. EGLN2

There are 181 ESTs and two full insert sequence cDNA clones (FIS) derived from this gene present in the human EGLN2 cluster. These assemble into a 2086 bp transcript with a 407 codon open reading frame. A single EST (BF690531) represents the skipping of exon 4. This would shift the reading frame and cause exon 5 to be translated in an altered reading frame for 29 codons before encountering a stop codon. The cDNA clone AK026863 retains intron 3, although other introns are spliced out, causing the premature termination of the reading frame 16 codons downstream of the normal exon 3 splice site. The mRNA AL133009 represents an unspliced transcript from the EGLN2 locus that initiates in intron 1 and reads through exons 2–5. However, the reading frame is not maintained in this transcript, which may represent genomic contamination of the cDNA library.

One hundred and fifteen mouse ESTs assemble into a 2089 bp contig representing *EGLN2*. As found to be the case for the human assembly, there are several ESTs clearly deriving from the *EGLN2* locus that exhibit differential splicing and intron retention. However, each of the alternatively processed transcripts are represented at less than 1% and none are expected to maintain an open reading frame. Both mouse and human *EGLN2* genes have consensus Kozak motifs (Kozak, 1996) at the predicted translation initiation sites and both utilize a single consensus polyadenylation signal for cleavage and polyadenylation.

3.1.3. EGLN3

Human *EGLN3* was identified from a cluster of 55 ESTs assembled into a contig of 2773 bp. The consensus sequence of the assembly is comprised of 327 bp of 5′ UTR, a 239 codon open reading frame and 1730 bp of 3′ UTR. Mouse *EGLN3* is represented by 49 ESTs, which were assembled into a 2655 bp contig. Both mouse and human assemblies show that two polyadenylation sites are used. In humans this shows a preference for the proximal site, while in mouse there is a preference for the distal site. In both species the two polyadenylation sites conform to the AATAAA consensus. The predicted translation initiation sites for both mouse and human conform to an adequate context (Kozak, 1996).

3.2. Genomic structure of human EGLN genes

The genomic structures of the human *EGLN* genes were determined through aligning the EST assemblies to draft human genome sequence, taking into account splice site consensus sequences to resolve alignment ambiguities.

EGLN1 was found to consist of five exons, which agrees with that previously reported (Dupuy et al., 2000). *EGLN2* and *EGLN3* genomic structures show the same pattern as *EGLN1*, having a large first exon and phase 0 intron. The second intron is also phase 0, intron 3 is phase 2 and intron 4 is phase 1. In each case, exon 5 has a stretch of coding sequence and subsequently runs directly into 3′ UTR without further splicing of the transcript.

3.3. Evolutionary relationship of the EGLN genes

In global alignment, human and mouse EGLN orthologue pairs share 84, 90 and 97% identity, respectively, for EGLN1, EGLN2 and EGLN3 at the amino acid level. *EGLN1* (*C1orf12*) has previously been reported to be the human orthologue of rat *Sm-20* (Dupuy et al., 2000). However, it is clear from multiple sequence alignment (Fig. 1) that rat SM-20 shows substantially greater homology with EGLN3 in aligned regions (>99% identity with mouse and 97% identity with human). The conclusion that EGLN3 is the true mouse and human orthologue of rat SM-20 is further supported by the observation of extended regions of homology in non-coding sequence between mouse/human *EGLN3* and *Sm-20*. No such similarity is observed between *Sm-20* and *EGLN1* or *EGLN2* (data not shown).

Rat SM-20 has been demonstrated experimentally to localize to the mitochondria. This localization is dependent on the first 25 amino acids of the protein (Lipscomb et al., 2001). It was also noted by Lipscomb et al. (2001) that the mitochondrial localization was not necessary for the role of SM-20 in the induction of apoptosis. Neither mouse or human *EGLN3* are predicted to encode amino acids corresponding to the first 116 amino acids of SM-20. While there is nucleotide homology between the 5′ coding sequence of *Sm-20* compared to human and mouse *EGLN3* transcripts, nucleotide substitutions and insertions/deletions between species contrive to disrupt the reading frame and introduce in-frame stop codons. In addition, the translation initiation site and start codon predicted to be used for *Sm-20* are absent in human *EGLN3*. Both human and mouse *EGLN3* genes are predicted to initiate translation at the equivalent of SM-20 amino acid 117. Since the mitochondrial targeting signal of SM-20 is located in the first 25 amino acids, it is unlikely that the mouse or human EGLN3 protein localizes to the mitochondria.

Human EGLN1 (but not EGLN2 or 3) and *C. elegans* Egl-9 proteins both contain an N-terminal MYND-type zinc finger motif (Pfam:PF01753) that shows 40% identity over 52 amino acids and conserves all of the residues critical for zinc binding (Fig. 1). This motif is conserved in *Fugu rubripes* EGLN1 (unpublished data) and also appears to be present in mouse EGLN1, although this sequence is incom-

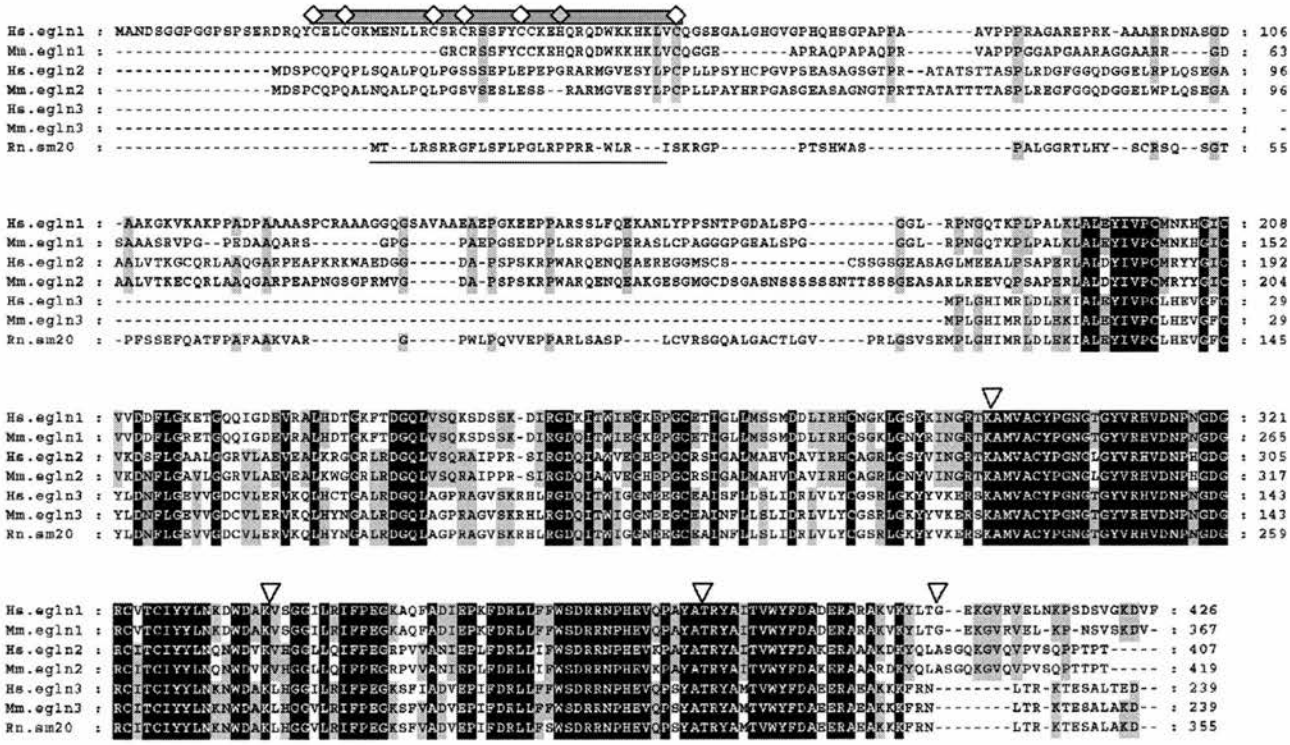


Fig. 1. Multiple sequence alignment of vertebrate EGLN protein, amino acid sequences. For background shading, black indicates 100% and grey 60% identity between aligned sequences. The horizontal grey bar indicates the extent of the MYND-type zinc finger in EGLN1 proteins and embedded diamonds indicate the conserved cysteine and histidine (grey diamond) residues. The mitochondrial targeting sequence of SM-20 is indicated by a horizontal black line. Triangles above the alignment show the position of splice sites relative to the encoded amino acid sequence for all of the human *EGLN* genes.

plete at the 5′ end (Fig. 1). The absence of this motif in SM-20 further supports the conclusion that SM-20 is orthologous to EGLN3 rather than EGLN1. The rat EST sequence AI510999 and overlapping cDNA fragments are strong candidates for the rat orthologue of *EGLN1* (data not shown).

3.4. Definition and phylogenetic distribution of the EGLN domain

A HMM of the conserved EGLN C-terminal domain (subsequently the EGLN domain) was generated from amino acid alignments (see Section 2.3) and used in conjunction with BLAST to search for other members of the EGLN domain containing a family of proteins. In *C. elegans* genomic sequence, the only family member identified is *egl-9*. In human genomic and EST sequences, all homology to the EGLN domain is accounted for by the *EGLN* 1–3 genes, *SCAND2*, and the *EGLN3* processed pseudogene located on chromosome 12 (Table 2). Considering the >90% sequence coverage of the human genome and depth of EST coverage for *EGLN* 1–3, it is likely that these three genes represent the total complement of this gene family in the human genome. A single EGLN domain-containing gene has been predicted from the *Drosophila* genomic sequence, the hypothetical 325 amino acid protein CG1114 (SPTR:Q9VN98). This *Drosophila* gene, as predicted, would encode a C-terminal EGLN homology domain and a less conserved N-terminal region.

The presence of a MYND-type zinc finger in Egl-9 and its conservation in one of the mammalian homologues suggested that the zinc finger–EGLN domain combination represents the ancestral form of the protein, which subsequently duplicated and diverged in the lineage leading to vertebrates. Although the predicted *Drosophila* protein CG1114 does not contain a zinc finger, the genomic sequence upstream (AE003603) has the potential to code for a MYND-type zinc finger, and is predicted as a separate, single exon gene (CG14665). Based on the cross-species evidence, it is likely that CG1114 and CG14665 actually represent a single gene, the *Drosophila* orthologue of Egl-9. This conclusion is supported by the finding that opposite end sequences of the cDNA clone LD24638 (AA820762 and AI455510) link the two predicted genes into a single transcript when *Drosophila* ESTs are assembled in the manner described in Section 2.2.

Aravind and Koonin (2001) have provided evidence that the conserved EGLN domain comprises a subfamily of the 2OG-Fe(II) oxygenase superfamily. This superfamily of proteins is widely distributed through eukaryotes, bacteria, archae and viruses. The *Streptomyces ansochromogenes* protein SANF (Q9KIT9) was also identified as a member of this protein superfamily (Aravind and Koonin, 2001). SANF was the highest scoring alignment with the EGLN domain HMMs (see Section 2.3), other than the identified EGLN family proteins (hmmsearch score of 18.0 versus 549.9–406.8 for EGLN family proteins). On this basis, the SANF amino acid sequence was used as an out group for phylogenetic analysis of the EGLN gene family (see Section 3.5).

3.5. Horizontal gene transfer?

Searching of the EST and genomic sequence databases identifies strong homologues of the EGLN subfamily in many vertebrates, *Drosophila melanogaster*, *C. elegans* and the ascidian *Ciona intestinalis* (AV675485 and AV680334), all of which are metazoan eukaryotes.

It is striking that given the apparent metazoan origin of this gene family, genes with the potential to encode proteins clearly of the EGLN subfamily were identified in the pathogenic bacteria *P. aeruginosa* and *Vibrio cholerae* (Fig. 2) (Aravind and Koonin, 2001). Close homologues of the *P. aeruginosa* EGLN-related gene (Q9I6I1) are also present in the unfinished bacterial genome sequences of *Pseudomonas syringae* and *Pseudomonas putida* (http://www.tigr.org/) (data not shown). However, the EGLN subfamily appears absent (based on HMM searching) from the genomes of the other fully and partially sequenced eubacteria, archae, yeast, plants and viruses.

This mosaic pattern of phylogenetic distribution is most readily explained by horizontal gene transfer from metazoans to bacteria (Aravind and Koonin, 2001), the other possibility being vertical inheritance and selective loss in many lineages, with specific retention in metazoans and some eubacteria. Phylogenetic analysis using *S. ansochromogenes SANF* amino acid sequence as an out group (Fig. 3) supports a vertical inheritance model rather than horizontal transfer between metazoans and eubacteria. *C. elegans* Egl-9 has previously been implicated in host–pathogen interactions with *P. aeruginosa* (Darby et al., 1999; Johnson and Liu, 2000). The presence of an unexpectedly close homologue in *P. aeruginosa* and other pathogenic bacteria suggests that the EGLN domain may be of significance to pathogenicity, although a mechanism for this remains unclear.

3.6. Conclusions

Human and mouse genomes encode three homologues of the *C. elegans egl-9* gene. Each of these homologues have a conserved gene structure suggesting duplication in the lineage leading to vertebrates. In the human genome, *EGLN1* maps to chromosome 1q42, *EGLN2* to 19q and *EGLN3* to 14q. *SCAND2* has previously been reported to map to chromosome 15 and there is an *EGLN3* processed pseudogene on chromosome 12p. Based on patterns of amino acid conservation and domain organization, *EGLN1* represents the ancestral form of this gene family in metazoans. Rat *Sm-20* is the orthologue of human and mouse *EGLN3*. The lack of potential mitochondrial targeting signals in *EGLN3* and the observation that mitochondrial targeting is not necessary for the role of SM-20 in caspase-mediated apoptosis suggest

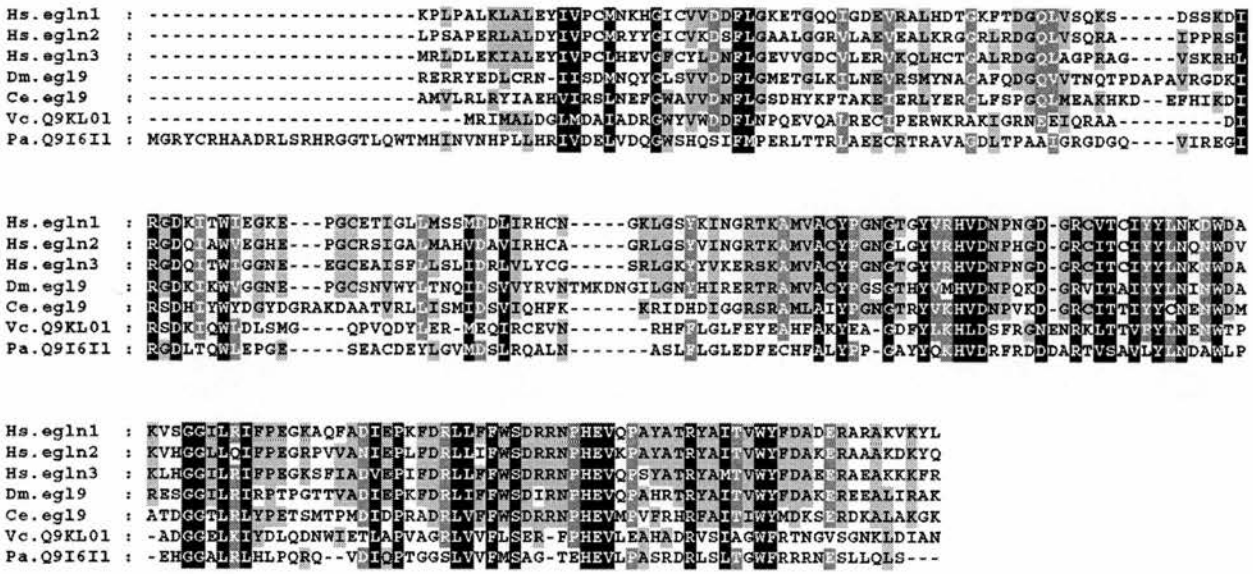


Fig. 2. Alignment of the conserved EGLN domain. The EGLN domains from the three human genes are shown aligned (Clustal) with *C. elegans* Egl-9 (Ce.egl9) and the *D. melanogaster* Egl-9 homologue CG1114 (Dm.egl9). Translations of the complete open reading frame from *V. cholerae* (Vc.Q9KL01) and *P. aeruginosa* (Pa.Q9I6I1) are also aligned.

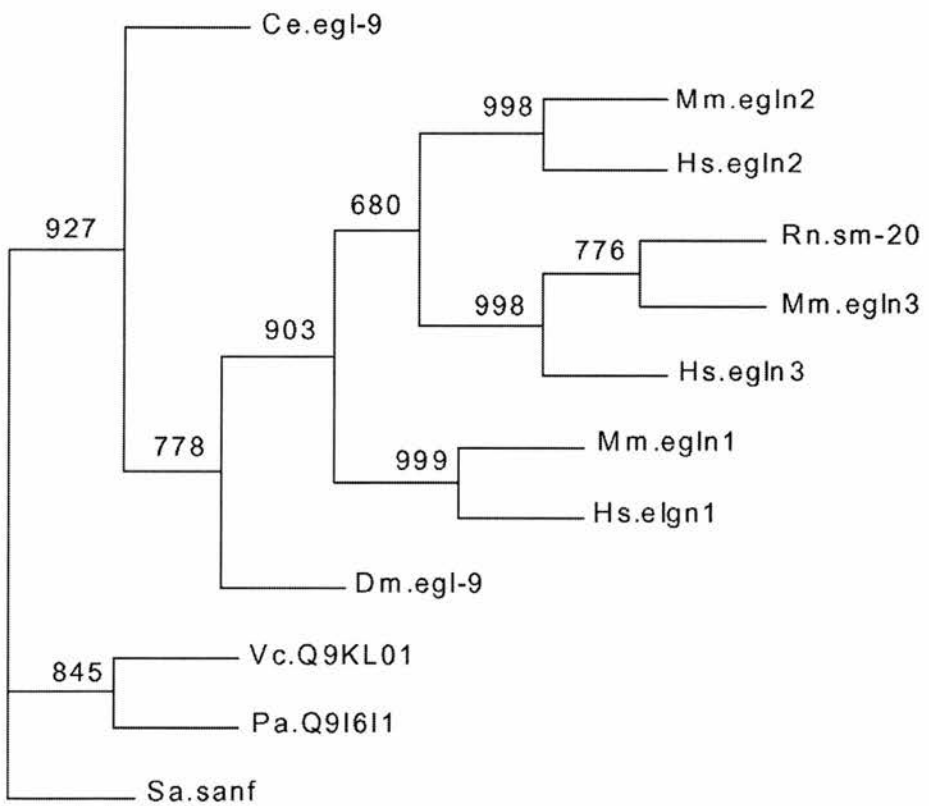


Fig. 3. Phenogram of the EGLN gene family. This unrooted tree is shown rerouted by the out group Sa.sanf (Q9KIT9). Numbers at branch forks indicate the number of times the group consisting of sequences to the right of that fork occurred among the trees out of 1000 replicates. See Section 2.4 for the method of tree construction and Section 3.4 for selection of Sa.sanf as an out group.

that mitochondrial targeting is not a general feature of the *EGLN* gene family and is likely to be a peculiarity of SM-20. The presence of *EGLN*-like genes in *Pseudomonas* and *V. cholerae* and their absence in other bacteria suggests highly lineage-specific retention of this gene family, although horizontal transfer cannot be ruled out. The known role of Egl-9 in the susceptibility of *C. elegans* to a *P. aeruginosa* toxin combined with the presence of an EGLN gene in this bacteria is an intriguing observation that provides new insight to host–pathogen interactions.

Acknowledgements

I would like to thank David J. Porteous, Rebecca S. Devon, Colin A.M. Semple and Andrea L. Bacon for critical reading of the manuscript and Fiona Brinkman for informative discussions.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Aravind, L., Koonin, E.V., 2001. The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. Genome Biol. 2, 7.7–7.8.

Darby, C., Cosma, C.L., Thomas, J.H., Manoil, C., 1999. Lethal paralysis of Caenorhabditis elegans by Pseudomonas aeruginosa. Proc. Natl. Acad. Sci. USA 96, 15202–15207.

Dupuy, D., Aubert, I., Duperat, V.G., Petit, J., Taine, L., Stef, M., Bloch, B., Arveiler, B., 2000. Mapping, characterization, and expression analysis of the SM-20 human homologue, c1orf12, and identification of a novel related gene, SCAND2. Genomics 69, 348–354.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8, 175–185.

Hurt, E.C., Schatz, G., 1987. A cytosolic protein contains a cryptic mitochondrial targeting signal. Nature 325, 499–503.

Johnson, C.D., Liu, L.X., 2000. Novel antimicrobial targets from combined pathogen and host genetics. Proc. Natl. Acad. Sci. USA 97, 958–959.

Kozak, M., 1996. Interpreting cDNA sequences: some insights from studies on translation. Mamm. Genome 7, 563–574.

Lipscomb, E.A., Sarmiere, P.D., Crowder, R.J., Freeman, R.S., 1999. Expression of the SM-20 gene promotes death in nerve growth factor-dependent sympathetic neurons. J. Neurochem. 73, 429–432.

Lipscomb, E.A., Sarmiere, P.D., Freeman, R.S., 2001. SM-20 is a novel mitochondrial protein that causes caspase-dependent cell death in nerve growth factor-dependent neurons. J. Biol. Chem. 276, 5085–5092.

Madden, S.L., Galella, E.A., Riley, D., Bertelsen, A.H., Beaudry, G.A., 1996. Induction of cell growth regulatory genes by p53. Cancer Res. 56, 5384–5390.

Moschella, M.C., Menzies, K., Tsao, L., Lieb, M.A., Kohtz, J.D., Kohtz, D.S., Taubman, M.B., 1999. SM-20 is a novel growth factor-responsive gene regulated during skeletal muscle development and differentiation. Gene Expr. 8, 59–66.

Mott, R., 1997. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput. Appl. Biosci. 13, 477–478.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Trent, C., Tsung, N., Horvitz, H.R., 1983. Egg-laying defective mutants of the nematode C. elegans. Genetics 104, 619–647.

Wax, S.D., Rosenfield, C.L., Taubman, M.B., 1994. Identification of a novel growth factor-responsive gene in vascular smooth muscle cells. J. Biol. Chem. 269, 13041–13047.

Genomic Sequence Analysis of *Fugu rubripes* CFTR and Flanking Genes in a 60 kb Region Conserving Synteny with 800 kb of Human Chromosome 7

Heather Davidson,[1,3] Martin S. Taylor,[1] Ann Doherty,[1] A. Christopher Boyd,[1] and David J. Porteous[1,2]

[1]*Medical Research Council Human Genetics Unit, Western General Hospital, Edinburgh EH4 2XU, UK;* [2]*Medical Genetics Section, Department of Medical Sciences, University of Edinburgh, Molecular Medicine Centre, Western General Hospital, Edinburgh EH4 2XU, UK*

To define control elements that regulate tissue-specific expression of the cystic fibrosis transmembrane regulator (CFTR), we have sequenced 60 kb of genomic DNA from the puffer fish *Fugu rubripes* (*Fugu*) that includes the CFTR gene. This region of the *Fugu* genome shows conservation of synteny with 800-kb sequence of the human genome encompassing the WNT2, CFTR, Z43555, and CBP90 genes. Additionally, the genomic structure of each gene is conserved. In a multiple sequence alignment of human, mouse, and *Fugu*, the putative WNT2 promoter sequence is shown to contain highly conserved elements that may be transcription factor or other regulatory binding sites. We have found two putative ankyrin repeat-containing genes that flank the CFTR gene. Overall sequence analysis suggests conservation of intron/exon boundaries between *Fugu* and human CFTR and revealed extensive homology between functional protein domains. However, the immediate 5′ regions of human and *Fugu* CFTR are highly divergent with few conserved sequences apart from those resembling diminished cAMP response elements (CRE) and CAAT box elements. Interestingly, the polymorphic polyT tract located upstream of exon 9 is present in human and *Fugu* but absent in mouse. Similarly, an intron 1 and intron 9 element common to human and *Fugu* is absent in mouse. The euryhaline killifish CFTR coding sequence is highly homologous to the *Fugu* sequence, suggesting that upregulation of CFTR in that species in response to salinity may be regulated transcriptionally.

[The sequence data described in this paper have been submitted to the GenBank data library under accession no. AJ271361, for the combined cosmids 159C9, 146H13, 6M15, and 145M20.]

The puffer fish *Fugu rubripes* (*Fugu*) has one of the smallest genomes (400 Mb) of all vertebrates, attributable to a compactness of introns, intergenic distances, and marked reduction of repetitive DNA sequences. Because the *Fugu* genome possesses a gene repertoire similar to that of other vertebrates, it is a valuable model for vertebrate gene analysis (Brenner et al. 1993).

The utility of *Fugu* genome analysis for the identification of candidate genes at disease loci and the demonstration of conserved synteny between species is well established (Aparicio et al. 1995; Armes et al. 1997; Baxendale et al. 1995; Gellner et al. 1999; Sandford et al. 1997; Schofield et al. 1997; Venkatesh et al. 1997; Yeo et al. 1997). It is hypothesized that the large evolutionary distance separating *Fugu* and mammals (about 350 million years) will have resulted in divergence of most sequences except for those of conserved functional or structural importance, such as coding and regulatory regions. Previous comparative sequence analyses between the genomes of *Fugu* and other species have identified sequences important for the control of gene expression (Aparicio et al. 1995; How et al. 1996; Marshall et al. 1994; Sandford et al. 1997; Venkatesh et al. 1996; Venkatesh et al. 1997). Using this approach, Gellner and Rowitch (Gellner et al. 1999; Rowitch et al. 1998) identified potential regulatory elements for *wnt1*, which encodes a protein expressed in the developing midbrain.

However, levels of conservation in noncoding regions can vary considerably between *Fugu* and other species. Comparison of the *Fugu* and human sequences in the Huntington's disease genomic region has not identified any conserved regulatory sequences (Baxendale et al. 1995). In contrast, in the region encompassing the WT1, Pax6 and RCN1 genes, there is significant noncoding homology between *Fugu* and human at the PAX6 locus, implying conservation of regulatory elements (Miles et al. 1998). However, human WT1 generates 16 protein isoforms (Miles et al. 1998), whereas the sequence data predicts that *Fugu* WT1 will only produce two isoforms.

Synteny is not always conserved between *Fugu* and

[3]Corresponding author.
E-MAIL H.Davidson@ed.ac.uk; FAX 44 131 651 1059.

10:1194–1203 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

mammalian genomes. In the Surfeit gene cluster, there is extensive rearrangement of genes between *Fugu* and human within a region of otherwise conserved synteny. This suggests that intrachromosomal rearrangements, probably inversions, have occurred during evolution (Gilley et al. 1997; Gilley et al. 1999). Despite these caveats, there is ample evidence that *Fugu* sequence analysis provides important information about regulatory elements, conserved synteny, splicing, and gene organization (Angrist 1998; Coutelle et al. 1998; Gellner et al. 1999; Gilley et al. 1999; How et al. 1996; Maheshwar et al. 1996; Marshall et al. 1994; Miles et al. 1998; Trower et al. 1996; Venkatesh et al. 1996; Yeo et al. 1997).

Our aim is to produce gene therapy vectors for cystic fibrosis (CF) that generate tissue-specific expression of the cystic fibrosis transmembrane conductance regulator (CFTR) at physiologic levels. We therefore require a thorough understanding of CFTR gene structure and regulation. The CFTR gene is known to be expressed in a tightly regulated fashion (Chou et al. 1991; Denamur et al. 1994; Matthews et al. 1996; Pittman et al. 1995; Trapnell et al. 1991; Yoshimura et al. 1991a,b). The control and enhancer regions for CFTR are not yet fully defined, although DNase I hypersensitive sites (DHSs) have been found at − 20.5 kb, − 79.5 kb, 185 + 10 kb within the first intron (Smith et al. 1995; Smith et al. 1996), and 4574 + 15.6 kb beyond the 3′ end of the gene (Nuthall et al. 1999). Putative cAMP response elements (CREs), Y-box, Ap-1, Sp-1, major and minor transcriptional start sites, and CAAT-like sequences have been identified by sequence analysis, 5′ RACE, mutational analysis, and electrophoretic mobility shift assays (Chou et al. 1991; Denamur et al. 1994; Imler et al. 1996; Matthews et al. 1996; Pittman et al. 1995; Vuillaumier et al. 1997; White et al. 1998; Yoshimura et al. 1991a).

To identify conserved elements that regulate CFTR expression, we have isolated and cloned *Fugu* CFTR. Using comparative sequence analysis, we identified regions of conserved synteny and putative exon/intron boundaries for the CFTR and flanking genes.

RESULTS AND DISCUSSION

Isolation of *Fugu* CFTR Cosmids

In search of sequences that control tissue-specific expression of CFTR, we identified and cloned *Fugu* CFTR with sufficient flanking sequence to encompass 5′ and 3′ control elements and neighboring genes. We screened a *Fugu* cosmid library using degenerate oligomers, compiled from an alignment of killifish, human, dogfish, mouse, and *Xenopus* coding sequences, as hybridization probes. Two separate hybridization experiments were performed and only those cosmids that gave a positive signal to both experiments were investigated further. We observed a weak signal in a common set of nine cosmids in both hybridization experiments and found them to be related by restriction analysis, Southern transfer, and hybridization to human CFTR probes. Preliminary sequence of one cosmid was highly homologous to exon 13 of killifish CFTR (data not shown). These data confirmed the isolation of a set of cosmids covering part of the *Fugu* CFTR region.

Initially, we subcloned one cosmid 159C9 into a pBluescript vector and made four libraries containing either AluI or Sau3a cosmid insert DNA. Larger EcoRI and PstI inserts were also generated for subcloning and sequencing. Finally, we used cosmid walking to fill gaps in the sequence. The consed sequence assembly package was used to contig the sequence data (Ewing et al. 1998a,b). We sequenced cosmid 159C9 entirely, which was shown to contain candidate *Fugu* CFTR exons 3–24 with extensive 3′ sequence (~29 kb). The insert ends of the other eight related cosmids were sequenced to find clones that would extend the sequence. Three cosmids, 146H13, 6M15, and 145M20, had additional sequence at the 5′ end and were used to complete the *Fugu* CFTR genomic sequence by cosmid walking. Cosmid end sequencing (data not shown), restriction analysis, Southern transfer, and hybridization experiments gave valuable information on the wider genomic organization of the region, complementing the CFTR sequencing data.

Conservation of Synteny

In this study, conservation of synteny in the CFTR region has been demonstrated to extend over more than 800 kb of human and 60 kb of *Fugu* genomic sequence. In particular, we found that orthologs of genes flanking human CFTR, WNT2 (Monkley et al. 1996) (which belongs to a large gene family encoding a group of secreted signalling molecules), Z43555 (a protein of unknown function), and CBP90 (Ohoka et al. 1998) (a brain-specific cortactin-binding protein) are conserved in order and orientation in *Fugu* (Fig. 1). Using exon 1 of WNT2 and exon 2 of CBP90 as anchor points between species, this corresponds to a 9.2-fold genomic compaction in *Fugu* relative to human.

Multiple sequence alignment to 192 nt upstream of the putative *Fugu* translational start site for WNT2 shows 48% identity of residues between human, mouse, and *Fugu*. In comparison, exon 1 alignment gives 58% identity. Short regions of locally high homology may represent binding sites for transcription factors or other regulatory elements. The presumed translational start site resembles the Kozak (Kozak 1996) consensus, and is highly conserved between each of the three species (10 nt perfect conservation). The WNT2 promoter appears to be a "housekeeping"

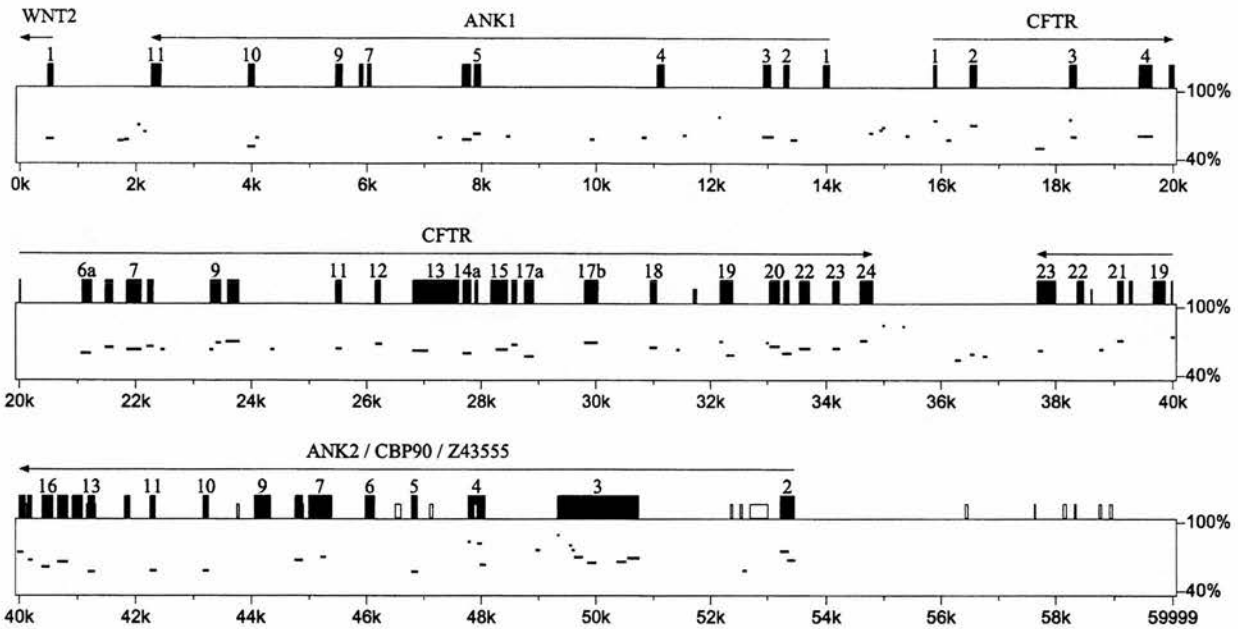


Figure 1 Genomic conservation and organization. Percentage identity plot (PIP) analysis of *Fugu* CFTR genomic region compared with the equivalent human region. Protein-coding exons are indicated by black rectangles. Gray and white boxes indicate 0.75 and 0.60 CpG:GpC ratios, respectively. Horizontal lines show blocks of homology between the *Fugu* and human sequences. Only blocks of homology occurring in the same relative order in both sequences are indicated by horizontal lines. Genes identified within the *Fugu* genomic sequence are annotated and transcriptional orientation indicated above the appropriate exons. Exons of genes are numbered where space permits.

type of promoter with elements conserved through evolution.

Interestingly, in *Fugu* we identified regions with ankyrin repeat homology on either side of the CFTR gene (Fig. 1). The ankyrin repeat homology 5′ to CFTR is described in mouse (designated MMU_Orf3; Ellsworth et al. 2000). Of the 13 predicted (Genscan) *Fugu* exons, 11 shared conservation with human and mouse as revealed by Dotter analysis. Complete intron/exon structures for the genes were determined by Genewise (E. Birney, unpubl.). The entire open reading frame (392 aa) of a hypothetical protein (ANK1 in *Fugu*) has been reconstructed. It contains four tandem ankyrin repeats and a Sterile Alpha Motif (SAM) domain (Pfam analysis), a similar domain arrangement to Tankyrase (a telomere-localized poly ADP-ribose polymerase; Smith et al. 1998).

There is a cluster 3′ to CFTR containing CBP90, ANK2, and Z43555 (Fig. 1). Genscan analysis of *Fugu* predicts six exons that share homology and correlate well with the exon/intron structure of the incomplete human coding region Z43555 (TREMBL accession no. 043388). A further two C-terminal exons, including an in-frame stop codon (Genscan), were predicted in *Fugu* and show conservation up to the stop codon in human sequence. ANK2 lies between CBP90 and Z43555 (Fig. 1) and contains 12 predicted exons (NIX; http://www.hgmp.mrc.ac.uk/NIX/) encoding putative ankyrin repeat motifs (Pfam analysis). The close proximity and consistency of orientation in *Fugu* and human, as well as an absence of potential polyadenylation signals (AAUAAA) within the region in *Fugu*, suggest that CBP90, ANK2, and Z43555 may all represent fragments of the same gene. It remains possible that one or more of the predicted exons are spurious. Further cDNA analysis is required to confirm the genetic structure. For supplementary information, see http://www.hgu.mrc.ac.uk/users/Heather.Davidson/Fugu.html. The comparison of relatively conserved coding sequences superimposed on diverged sequence background in *Fugu*, human, and mouse demonstrates the utility of *Fugu*: mammal comparison for the identification of putative novel genes.

Fugu Versus Mammalian CFTR

At the predicted amino acid level, a global alignment of *Fugu* and human CFTR demonstrates 58% identity and 75% similarity. In comparing human and *Fugu* functional CFTR domains, NBD1 (nucleotide-binding domain) exhibits 71% identity and 87% similarity, NBD2 58% identity and 73% similarity, the R (regulatory) domain 46% identity and 63% similarity, and MSD1 (membrane-spanning domain) 61% identity and 76% similarity, and MSD2 59% identity and 73% similarity. The fourth extracellular loop encoded within exon 15 and three regions of the R domain dis-

play a relatively high level of sequence divergence (Fig. 2).

Within the R domain of human CFTR, there are nine dibasic consensus phosphorylation sites for cAMP-dependent phosphorylation, a process critical for the regulation of CFTR Cl channel activity (Riordan et al. 1989; Xiu-Bao et al. 1993). All but two of these sites are conserved in *Fugu* CFTR (Fig. 2). Of the remaining sites, serine 700 is converted to a monobasic consensus phosphorylation site, while threonine 788 is abolished in *Fugu* CFTR. Interestingly, serine 670, a monobasic consensus phosphorylation site in human CFTR, is dibasic in *Fugu*, killifish, and mouse. Of the two N-glycosylation sites within the fourth extracellular loop of human CFTR, only the C-terminal site is predicted (Bause 1983) to be glycosylated in *Fugu* and killifish (Fig. 2).

The alignment of human, mouse, and *Fugu* CFTR genomic sequences as guided by the predicted amino acid sequences suggests that the relative position and coding phase of exons is conserved. Such conservation of exon position and phase suggests that equivalents of all known human splice forms could be generated from the *Fugu* ortholog. For each exon, the 40-nt flanking splice donor and acceptor sites from both *Fugu* and human CFTR were isolated and compared directly. Although conservation was calculated for both intron and exon components of each alignment, no correlated divergence from core splice consensus sequences was found. Interestingly, the *Fugu* intron 9 splice acceptor region contains 5′-TTTTTTTT-3′, possibly equivalent to the polymorphic polyT tract that affects the variable in-frame skipping of exon 9 of human CFTR (Chu et al. 1991, 1993). However, this polyT tract is absent in mouse (Ellsworth et al. 2000).

The 3′ splice site at the intron 9/exon 10 junction of CFTR is a good match with the consensus YYY_n-NYAG/– (Moore 2000) and is conserved between human, mouse, and *Fugu* (Fig. 3A). Within the intron and nearly equidistant (17–18 nt), in both human and *Fugu*, from the splice junction, there is a block of conservation (13–14 nt) whose position and resemblance to the consensus makes it a good candidate to be the splice branch site. The equivalent mouse sequence also contains consensus splice branch point homology, but the block of homology shared between human and *Fugu* is specifically absent in the mouse (Fig. 3A). The level of conservation and consistency of position makes this an intriguing observation, though its significance is unclear.

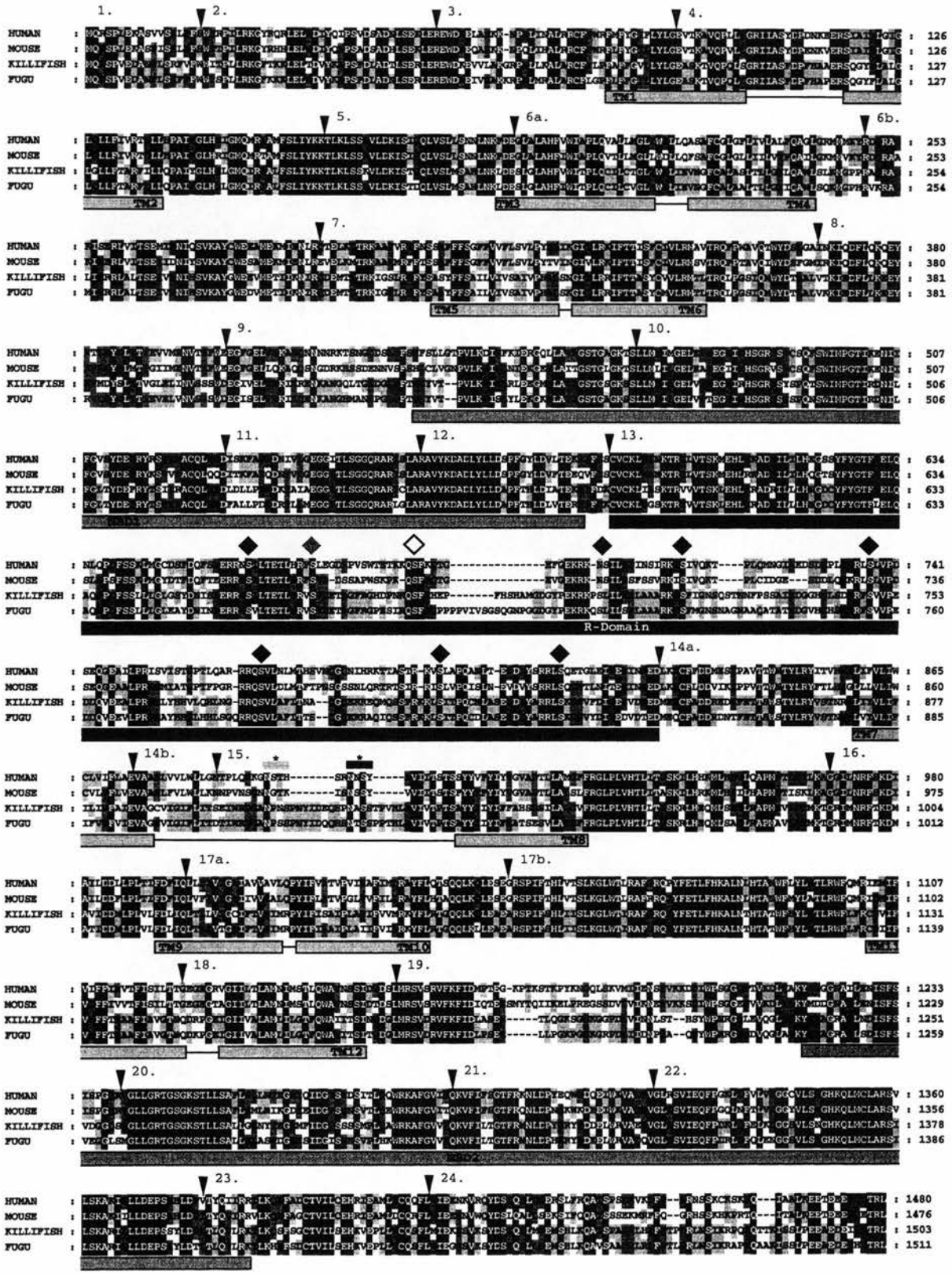
CFTR Transcriptional Regulation

Some transcription control elements of the CFTR gene are currently defined solely by homology to short consensus sequences. Our interest is to identify these elements which are functional and incorporate them into genomic context vectors for CF gene therapy. Little is currently known about the general sequence conservation and particular features of housekeeping gene promoters conserved between distantly related vertebrates.

Multiple sequence alignment (ClustalW; Thompson et al. 1994) of sequences directly upstream of CFTR coding sequence (data not shown) demonstrates good sequence conservation between mammals (Vuillaumier et al. 1997), but not between mammals and fish. Within the CFTR core promoter region of the aligned mammalian species, previously proposed regulatory sequence elements are generally poorly conserved (Fig. 3B). The Sp-1 sites, Ap-1 sites, and putative negative regulators of human CFTR (Fig. 3C; Chou et al. 1991; Denamur et al. 1994) are not highly conserved between salmon, *Fugu*, killifish, primate, bovine, or rabbit, questioning their importance in core promoter activity and tissue-specific expression (Fig. 3B). We have found no Sp-1 or Ap-1 sites within the *Fugu* CFTR promoter region, as has been described for the *Fugu* Surfeit family of genes (Gilley et al. 1997, 1999), which have housekeeping promoters similar to CFTR. Moreover, a TATA box at –545 bp in the CFTR promoter region is unique to *Fugu* CFTR. These sequence alterations may represent genuine differences in housekeeping gene regulation between mammals and fish.

Blocks of conserved promoter sequence in CFTR from mammals suggest functional significance, but the equivalent *Fugu* regions are again devoid of recognizable homology to these sequences (data not shown). However, homology to a CRE (TGACGTCA; Matthews et al. 1996) is found in *Fugu* at position –282 bp (TGACGT), while slight homology to an inverted Y box (consensus CWGATTGGYCYA) is found at position –344 bp (CAGATTCTATAT). The inverted and imperfect CAAT box (AATTGGAAGCAAAT) with conserved residues TTGGAAGCART (found in human, two primates, cow, and rabbit) is not completely conserved in *Fugu*, although at position –174 bp, there is the well-conserved GAGGAGAAGCAAGA motif. In mammalian species, the CRE and Y box normally overlap, whereas in *Fugu*, they do not. These data highlight the highly divergent nature of the *Fugu* genome and the lack of conserved regulatory elements between human and *Fugu* in the CFTR promoter region.

In the wider genomic context, no substantial blocks of conserved sequence (phylogenetic footprints) were identified in proximity to CFTR. We found no substantial regions of homology when we performed percentage identity plot (PIP) pairwise comparisons between *Fugu* and mouse and *Fugu* and human genomic sequences (Fig. 1; http://www.hgu.mrc.ac.uk/users/Heather.Davidson/Fugu.html). PIP analysis, however, supports the human exon predictions for the two an-


1.
2.
3.
4.
HUMAN
MOUSE
KILLIFISH
FUGU
126
126
127
127
TM1
5.
6a.
6b.
253
253
254
254
TM2
TM3
TM4
7.
8.
380
380
381
381
TM5
TM6
9.
10.
507
507
506
506
11.
12.
13.
634
634
633
633
741
736
753
760
R-Domain
14a.
865
860
877
885
TM7
14b.
15.
16.
980
975
1004
1012
TM8
17a.
17b.
1107
1102
1131
1139
TM9
TM10
TM11
18.
19.
1233
1229
1251
1259
TM12
20.
21.
22.
1360
1356
1378
1386
RHD2
23.
24.
1480
1476
1503
1511

kyrin repeat-containing proteins, Z43555 and CBP90. In addition to the predicted exons of CFTR and flanking genes, PIP analysis suggested several other regions of potential conservation (unannotated horizontal lines in Fig. 1); of these, all in the CFTR region can be attributable to regions of low compositional complexity, producing multiple spurious hits when the PIP chaining and single coverage models are not used (see Methods). Of particular interest is the absence of detectable homology in regions previously identified as containing DHSs (Nuthall et al. 1999; Smith et al. 1995, 1996). The DHSs at −79.5 and −20 kb show weak correlation with CFTR expression in both cultured epithelial cell lines and primary duct epithelial cells, and we found no evidence for their conservation in *Fugu* (Smith et al. 1995). The DHS at −79.5 in human will place it in the middle of the ankyrin repeat-containing protein, which has yet to be proven to be functional, and therefore this DHS site may not be involved in CFTR regulation.

The DHS in intron 1 (position 181 + 10 kb; Smith et al. 1996) correlates well with the levels of CFTR expression in these cell lines (Smith et al. 1995). Inclusion of human CFTR intron 1 has also been directly shown to confer transcriptional upregulation in a controlled reporter system (Mogayzel and Ashlock, 2000). Interestingly, we have found a complex element in *Fugu* intron 1 which shows noncoding homology between the human and *Fugu* species. This intron 1 element is too short to be picked up by PIP, but was found using dot matrix methods (data not shown; http://www.hgu.mrc.ac.uk/users/Heather.Davidson/Fugu.html). The element contains palindromic and direct repeat features (Fig. 3D). However, in the orthologous mouse genomic sequence (Ellsworth et al. 2000), there is an insertion in this element which casts doubt on its significance. It overlaps a known cluster of DNase I hypersensitive sites and DNase I footprints in human CFTR that correlate well with CFTR expression. Specifically, we have found a 17-bp element conserved between human intron 1 element A (HA) and *Fugu* intron 1 element A (FA). FA is conserved at 13/17 nucleotides between human and *Fugu* (Fig. 3D), corresponding to coordinates 181 + 9.727 kb and 181 + 135 bp, respectively. The orientation of FA is conserved with respect to CFTR transcription. Interestingly, within the *Fugu* intron 1, there is a second, inverted copy (*Fugu* intron 1 element B [FB]) 80 bp downstream. FB in *Fugu* shares 11/17 nucleotides with FA and 14/17 nucleotides with the human HA (Fig. 3D). However, no human equivalent of FB was identified.

At the 3′ end of the human CFTR gene, there is a DHS at position 4574 + 15.6 kb that regulates tissue-specific expression of CFTR (Nuthall et al. 1999). There is no evidence of conservation of this site in *Fugu*. This DHS and/or the other DHSs (at 4574 + 5.4 − 7.4 kb; Nuthall et al. 1999), which do not regulate CFTR expression, might instead regulate the expression of the Z43555 and CBP90 (two genes located 48 kb and 160 kb 3′ from the end of the CFTR gene).

Killifish CFTR Regulation

The euryhaline killifish (Singer et al. 1998) adapts rapidly to extreme changes in salinity. Exposure of freshwater-adapted killifish to seawater increases expression of killifish CFTR, implying a role for CFTR in salinity adaptation. Despite *Fugu* not being similarly adapted, its CFTR protein sequence is highly homologous (84% identity and 91% similarity) to that of killifish. Therefore, the CFTR-mediated freshwater adaptation of killifish is likely to be solely explained by transcriptional upregulation rather than fundamental differences in the property of the channel.

Summary

Our study has shown conservation of synteny and orientation between *Fugu* and human over a large, multigenic region. All genes identified appear to be capable of functioning in both species. Like CFTR, WNT2 has a housekeeping promoter. However, the WNT2 promoter has conserved elements between human, mouse, and *Fugu*, whereas there is little conservation in the promoter of *Fugu* CFTR. This suggests that regulation of WNT2 is far more conserved that that of CFTR. The data are consistent with complete conservation of CFTR exon/intron boundaries between *Fugu* and human, and there is extensive homology between functional domains. In noncoding regions of the CFTR gene, human and *Fugu* are highly divergent, and apart from a diminished CRE and CAAT box, the putative promoter regions are devoid of conserved regulatory domains. The inclusion of the mouse orthologous genomic sequence in the intron 1 comparison casts doubt on the validity of the identified conserved element identified in human and *Fugu* even though it is near a previously described DHS (Nuthall et al. 1999). The element's importance must be determined empirically to evaluate its functional significance. However, the additional finding of the polyT tract and the intron 9 element, both of which are present in human and

Figure 2 Sequence homology and conservation of exon/intron boundaries. Alignment of human, mouse, killifish, and *Fugu* CFTR. The triangles mark approximate intron/exon boundaries. Domains of CFTR are as indicated: NBD = nucleotide-binding domain; R-domain = regulatory domain; TM = transmembrane segment. Horizontal lines linking transmembrane segments represent extracellular loops. Diamonds above the phosphorylated residues predict dibasic cAMP-dependent phosphorylation sites. Short horizontal bars marked with asterisks above exon 15 indicate glycosylation sites. In the background shading at each position, phosphorylation and glycosylation sites: black indicates perfect homology, dark grey three out of four, and light grey two out of four residues matching.

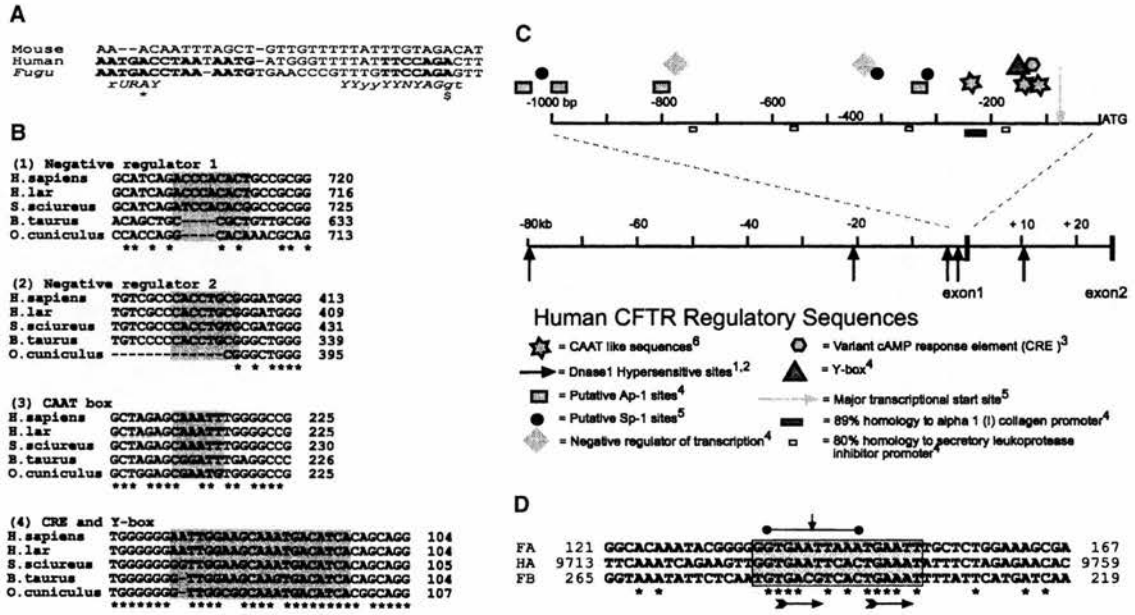


Figure 3 Putative human CFTR regulatory domains. Relative positions of proposed human CFTR transcription regulatory sequences and their conservation between mammals. (*A*) CFTR Intron 9/exon 10 3′ splice site aligned in mouse, human, and *Fugu*. Bold type indicates blocks of sequence conserved between human and *Fugu*, but not in mouse. Italic sequence shows consensus branch-point and 3′ splice site sequences. The proposed branch-point *A* is marked with *. $ indicates the first nucleotide of exon 10. Capitalized letters of consensus sequences summarize positions where both *Fugu* and human match the consensus. (*B*) Alignments 1 to 4 illustrate the conservation of selected CFTR promoter elements between human (Homo sapiens), gibbon (Hylobates lar), monkey (Saimiri sciureus), cow (Bos taurus) and rabbit (Oryctolagus cuniculus) at coordinates relative to the start of translation of human CFTR. Asterisks indicated nucleotides conserved between all aligned species. Background shading marks the extent of proposed elements. (*C*) The relative position of sequence motifs implicated in CFTR regulation, DNase I hypersensitive sites, and exons one and two are indicated with respect to the translational start site (ATG). Annotated elements are as defined in: [1](Smith et al. 1995), [2](Smith et al. 1996), [3](Matthews et al. 1996), [4](Chou et al. 1991), [5](Yoshimura et al. 1991a), [6](Pittman et al. 1995). (*D*) FA and FB *Fugu* CFTR intron 1 elements are shown in alignment with the conserved human HA element. Asterisks indicate nucleotides conserved in all three elements. The 10-bp perfect palindromes of FB and HA are indicated by a horizontal bar, and the axis of symmetry is indicated by a vertical arrow. Horizontal arrows indicate partially conserved direct repeats in each element. The coordinates indicate distance from the 3′ end of CFTR exon 1. Background shading marks the extent of the proposed element.

Fugu but absent in mouse, suggests that CFTR regulation in the mouse lineage has evolved eccentrically. There may be a correlation between these observations and the known differences between mouse and human in CFTR expression patterns (Manson et al. 1997) and mutant phenotypes (Davidson et al. 1995).

This and other studies (Ellsworth et al. 2000) of CFTR genomic structure and its conservation between species will inform our overall strategy of producing minimal genomic context vectors that provide regulated tissue-specific expression for CF gene therapy.

METHODS

Southern Blots and Hybridizations

The *Fugu* genomic cosmid library (coverage eightfold) was obtained from the Medical Research Council Human Genome Mapping Project Resource Centre. To test the degenerate oligomers, hybridization experiments at various temperatures and washing conditions were performed on a human CFTR P1 artificial chromosome (Ioannou et al. 1994), whose sequence was expected to be as divergent from the degenerate oligomer as that of a *Fugu* CFTR sequence. Conditions for the 53-mer degenerate oligomer (hybridization experiment 1) were optimized to be: hybridization at 60°C with three washes in 4× SSC (1× saline sodium citrate [SSC] is 0.15 M NaCl, 0.015 M sodium citrate), 0.1% sodium dodecyl sulphate (SDS), one at room temperature followed by two at 60°C. For the mixed oligomer hybridization (hybridization experiment 2), the conditions were optimized to be: hybridization at 48°C followed by three washes in 4× SSC, 0.1% SDS at room temperature, and one at 37°C. The Southern transfer blots of the restriction digested *Fugu* cosmid DNA were hybridized at 58°C and washed at 68°C three times in 2× SSC, 0.1% SDS.

The oligomers used in the hybridization were:

Exon 10, 53 bp

ATGATGATIITGGGIGAIITGGIGCCATCAGAIGGTAAAATIAII CACAGTGG

Exon 9, 24 bp

GCTGGATCTACIGGITCIGGIAAG

Exon 10, 30 bp

CCACTGTGIITIATTTTACCITCTGAACCG

Exon 11, 20 bp killifish

CTTGCTCTTTGACCCCCACT

Exon 10, 29 bp killifish

CCATCAGAGGGTAAAATCAGACACAGTGG

Cloning and Sequencing

The library was cloned into pBluescript vector, and subclones were sequenced with KS and SK primers (Stratagene Inc.) and ABI dye terminator chemistry using an ABI377 automated DNA sequencer. Sequences were assembled with the consed sequence assembly program (http://www.genome.washington.edu; Ewing et al. 1998a,b). Sequencing of cosmids with custom primers was used to close gaps and complete both strands. The sequence coverage for the *Fugu* CFTR genomic region is at least two high-quality sequence runs in both directions, coding regions being more extensively sequenced. In two noncoding regions, both situated between exon 18 and 19 of CFTR, short GC-rich tracts caused high-quality sequencing to be obtained in one direction only, despite several attempts. Another region we sequenced in one direction only is located at 21458 bp 3′ to the end of the coding sequence of CFTR between Z43555 and CBP90 (Fig. 3).

Comparative Sequence Analysis

Nucleotide sequences (Washington University Genome Sequencing Center) for human-derived bacterial artificial chromosomes were used to assemble 864 kb (from 7q31–32) of contiguous human genomic sequence, including and flanking the CFTR transcriptional unit (GenBank database accession nos. AC000061, AC000111, AC002465, AC003045, AC004240, and AC002431). Assembly was performed by iterative Fasta (Pearson et al. 1988) searches to determine overlapping coordinates, and final assembly was achieved using in-house software.

For human genomic sequence, a sliding window method of sequence fragmentation was used to generate an artificial contig of sequential 100-kb fragments with 50-kb overlaps. Each human fragment and the *Fugu* sequence was analyzed using programs for feature prediction and homology searching of appropriate public databases, coordinated through the NIX interface (http://www.hgmp.mrc.ac.uk/NIX/). Putative exons and genes were compared at the nucleotide and encoded amino acid level to known genes and the dbEST database using the BLAST (Altschul et al. 1997), Fasta (Pearson and Lipman, 1988), and HMMer (Eddy 1996) homology search tools. Exact coordinates of coding sequence and intron/exon boundaries were predicted using the Wise2 package (Birney et al. 1996). Homology templates used in Wise2 splice site prediction were killifish CFTR, rat CBP90, human WNT2, and Z43555 amino acid sequences (TREMBL accession nos. O73677, O88864, P09544, and O43388, respectively).

We performed PIP (PipMaker http://globin.cse.psu.edu/pipmaker/) pairwise comparison between *Fugu* and human and *Fugu* and mouse genomic sequences. For PIP analysis, low-complexity regions of the *Fugu* sequence were masked using RepeatMasker (A. Smit and P. Green, unpubl.). PipMaker default settings were adjusted so that match = 1, transition = −0.6, transversion = −0.8, and alignment cutoff = 18. The comparisons were only made in the forward strand. Chaining and single-coverage models of alignment distribution were used to reduce spurious matches (http://globin.cse.psu.edu/pipmaker/pip-instr2.html).

The multiple nucleotide sequence alignment of mammalian CFTR promoters (data not shown) was performed using ClustalW (Thompson et al. 1994) with default parameters for nucleic acid alignment. Nucleotide sequences directly upstream of CFTR coding sequence were too divergent between fish and mammals to construct an informative multiple sequence alignment (data not shown). The alignment of human, mouse, killifish, and *Fugu* CFTR amino acid sequences (Fig. 2) was achieved using ClustalW (Thompson et al. 1994). Pairwise comparisons between human and *Fugu* CFTR at the nucleotide and amino acid levels were carried out using ALIGN (Pearson and Lipman, 1988) default parameters. The GenBank accession codes for the various CFTR protein or DNA sequences were: human protein P13569, mouse protein P26361, killifish protein AAC41271, human promoter AC000111, mouse promoter L04873, cow promoter X95926, rat promoter X95927, squirrel monkey promoter X95928, salmon promoter AF155237, killifish promoter AF000271, crab-eating macaque gene X95929, gibbon gene X95930, rabbit gene X95931, and *Xenopus* promoter X65256.

Fugu and human genomic sequences were directly compared after fragmentation with a window size of 10 kb and 1 kb using Fasta and Lalign (Pearson and Lipman 1988) with known coding sequence masked. *Fugu* genomic sequence was further fragmented into sequential 100-bp blocks with 50-bp overlap and compared to human sequences using Fasta. Further analyses used a pairwise dot-matrix analysis performed with the Dotter program (Sonnhammer and Durbin 1995) at window sizes of 25 and 45 nucleotides.

ACKNOWLEDGMENTS

We thank Prof. Nicholas Hastie and Dr. David Sheppard for helpful comments on the manuscript, Stewart McKay and Agnes Gallagher for DNA sequencing, Webb Miller for help and advice with the optimization of the PIP analysis, and the United Kingdom Human Genome Mapping Project Resource Centre for supplying the *Fugu* cosmids. This work was supported by the UK Medical Research Council.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Angrist, M. 1998. Less is more: Compact genomes pay dividends. *Genome Res.* **8:** 683–685.

Aparicio, S., Morrison, A., Gould, A., Gilthorpe, J., Chaudhuri, C., Rigby, P., Krumlauf, R., and Brenner, S. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. *Proc. Natl. Acad. Sci.* **92:** 1684–1688.

Armes, N., Gilley, J., and Fried, M. 1997. The comparative genomic structure and sequence of the surfeit gene homologs in the puffer fish *Fugu rubripes* and their association with CpG-rich islands. *Genome Res.* **7:** 1138–1152.

Bause, E. 1983. Structural requirements of N-glycosylation of proteins. Studies with proline peptides as conformational probes. *Biochem. J.* **209:** 331–336.

Baxendale, S., Abdulla, S., Elgar, G., Buck, D., Berks, M., Micklem, G., Durbin, R., Bates, G., Brenner, S., and Beck, S. 1995. Comparative sequence analysis of the human and pufferfish Huntington's disease genes. *Nat. Genet.* **10:** 67–76.

Birney, E., Thompson, J.D., and Gibson, T.J. 1996. PairWise and SearchWise: Finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.* **24:** 2730–2739.

Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B., and Aparicio, S. 1993. Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* **366:** 265–268.

Chou, J.L., Rozmahel, R., and Tsui, L.C. 1991. Characterization of the promoter region of the cystic fibrosis transmembrane conductance regulator gene. *J. Biol. Chem.* **266:** 24471–24476.

Chu, C.S., Trapnell, B.C., Curristin, S., Cutting, G.R., and Crystal, R.G. 1993. Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA. *Nat. Genet.* **3:** 151–156.

Chu, C.S., Trapnell, B., Murtagh, J., Moss, J., Dalemans, W., Jallat, S., Mercenier, A., Pavirani, A., Lecocq, J.P., Cutting, G.R., et al. 1991. Variable deletion of exon 9 coding sequences in cystic fibrosis transmembrane conductance regulator gene mRNA transcripts in normal bronchial epithelium. *EMBO J.* **10:** 1355–1363.

Coutelle, O., Nyakatura, G., Taudien, S., Elgar, G., Brenner, S., Platzer, M., Drescher, B., Jouet, M., Kenwrick, S., and Rosenthal, A. 1998. The neural cell adhesion molecule L1: Genomic organisation and differential splicing is conserved between man and the pufferfish *Fugu*. *Gene* **208:** 7–15.

Davidson, D.J., Dorin, J.R., McLachlan, G., Ranaldi, V., Lamb, D., Doherty, C., Govan, J., and Porteous, D.J. 1995. Lung disease in the cystic fibrosis mouse exposed to bacterial pathogens. *Nat. Genet.* **9:** 351–357.

Denamur, E. and Chehab, F.F. 1994. Analysis of the mouse and rat CFTR promoter regions. *Hum. Mol. Genet.* **3:** 1089–1094.

Eddy, S.R. 1996. Hidden Markov models. *Curr. Opin. Struct. Biol.* **6:** 361–365.

Ellsworth, R.E., Jamison, D.C., Touchman, J.W., Chissoe, S.L., Braden, M.V., Bouffard, G.G., Dietrich, N.L., Beckstrom-Sternberg, S.M., Iyer, L.M., Weintraub, L.A. et al. 2000. Comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane conductance regulator genes *Proc. Natl. Acad. Sci.* **97:** 1172–1177.

Ewing, B. and Green, P. 1998a. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8:** 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998b. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Gellner, K. and Brenner, S. 1999. Analysis of 148 kb of genomic DNA around the wnt1 locus of *Fugu rubripes*. *Genome Res.* **9:** 251–258.

Gilley, J., Armes, N., and Fried, M. 1997. *Fugu* genome is not a good mammalian model. *Nature* **385:** 305–306.

Gilley, J. and Fried, M. 1999. Extensive gene order differences within regions of conserved synteny between the fugu and human genomes: implications for chromosomal evolution and the cloning of disease genes. *Hum. Mol. Genet.* **8:** 1313–1320.

How, G.F., Venkatesh, B., and Brenner, S. 1996. Conserved linkage between the puffer fish (*Fugu rubripes*) and human genes for platelet-derived growth factor receptor and macrophage colony-stimulating factor receptor. *Genome Res.* **6:** 1185–1191.

Imler, J.L., Dupuit, F., Chartier, C., Accart, N., Dieterle, A., Schultz, H., Puchelle, E., and Pavirani, A. 1996. Targeting cell-specific gene expression with an adenovirus vector containing the lacZ gene under the control of the CFTR promoter. *Gene Ther.* **3:** 49–58.

Ioannou, P.A., Amemiya, C.T., Garnes, J., Kroisel, P.M., Shizuya, H., Chen, C., Batzer, M.A., and Dejong, P.J. 1994. A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nat. Genet.* **6:** 84–89.

Kozak, M. 1996. Interpreting cDNA sequences: some insights from studies on translation. *Mamm. Genome* **7:** 563–574.

Maheshwar, M.M., Sandford, R., Nellist, M., Cheadle, J.P., Sgotto, B., Vaudin, M., and Sampson, J.R. 1996. Comparative analysis and genomic structure of the tuberous sclerosis 2 (TSC2) gene in human and pufferfish [published erratum appears in *Hum. Mol. Genet.* 1996 Apr;**5**(4):562]. *Hum. Mol. Genet.* **5:** 131–137.

Manson, A.L., Trezise, A.E., MacVinish, L.J., Kasschau, K.D., Birchall, N., Episkopou, V., Vassaux, G., Evans, M.J., Colledge, W.H., Cuthbert, A.W. et al. 1997. Complementation of null CF mice with a human CFTR YAC transgene. *EMBO J.* **16:** 4238–4249.

Marshall, H., Studer, M., Popperl, H., Aparicio, S., Kuroiwa, A., Brenner, S., and Krumlauf, R. 1994. A conserved retinoic acid response element required for early expression of the homeobox gene Hoxb-1. *Nature* **370:** 567–571.

Matthews, R.P. and McKnight, G.S. 1996. Characterization of the cAMP response element of the cystic fibrosis transmembrane conductance regulator gene promoter. *J. Biol. Chem.* **271:** 31869–31877.

Miles, C., Elgar, G., Coles, E., Kleinjan, D.J., van Heyningen, H.V., and Hastie, N. 1998. Complete sequencing of the *Fugu* WAGR region from WT1 to PAX6: Dramatic compaction and conservation of synteny with human chromosome 11p13. *Proc. Natl. Acad. Sci.* **95:** 13068–13072.

Monkley, S.J., Delaney, S.J., Pennisi, D.J., Christiansen, J.H., and Wainwright, B.J. 1996. Targeted disruption of the Wnt2 gene results in placentation defects. *Development* **122:** 3343–3353.

Moore, M.J. 2000. Intron recognition comes of Age. *Nat. Struct. Biol.* **7:** 14–16.

Mogayzel, P.J. Jr. and Ashlock, M.A. 2000. CFTR intron 1 increases luciferase expression driven by CFTR 5′-flanking DNA in a yeast artificial chromosome. *Genomics* **64:** 211–215.

Nuthall, H.N., Moulin, D.S., Huxley, C., and Harris, A. 1999. Analysis of DNase-I-hypersensitive sites at the 3′ end of the cystic fibrosis transmembrane conductance regulator gene (CFTR). *Biochem. J.* **341:** 601–611.

Ohoka, Y. and Takai, Y. 1998. Isolation and characterization of cortactin isoforms and a novel cortactin-binding protein, CBP90. *Genes Cells* **3:** 603–612.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Pittman, N., Shue, G., Leleiko, N.S., and Walsh, M.J. 1995. Transcription of cystic fibrosis transmembrane conductance regulator requires a CCAAT-like element for both basal and cAMP-mediated regulation. *J. Biol. Chem.* **270:** 28848–28857.

Riordan, J.R., Rommens, J.M., Kerem, B.S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S. , Plavsic, N., Chou, J.L. et al. 1989. Identification of the cystic-fibrosis gene-cloning and characterization of complementary-DNA. *Science* **245:** 1066–1072.

Rowitch, D.H., Echelard, Y., Danielian, P.S., Gellner, K., Brenner, S., and McMahon, A.P. 1998. Identification of an evolutionarily conserved 110 base-pair cis-acting regulatory sequence that governs Wnt-1 expression in the murine neural plate. *Development* **125:** 2735–2746.

Sandford, R., Sgotto, B., Aparicio, S., Brenner, S., Vaudin, M., Wilson, R.K., Chissoe, S., Pepin, K., Bateman, A., Chothia, C. et al. 1997. Comparative analysis of the polycystic kidney disease 1 (PKD1) gene reveals an integral membrane glycoprotein with multiple evolutionary conserved domains. *Hum. Mol. Genet.* **6:** 1483–1489.

Schofield, J.P., Elgar, G., Greystrong, J., Lye, G., Deadman, R., Micklem, G., King, A., Brenner, S., and Vaudin, M. 1997. Regions of human chromosome 2 (2q32-q35) and mouse chromosome 1 show synteny with the pufferfish genome (*Fugu rubripes*). *Genomics* **45:** 158–167.

Singer, T.D., Tucker, S.J., Marshall, W.S., and Higgins, C.F. 1998. A divergent CFTR homologue: Highly regulated salt transport in the euryhaline teleost F. heteroclitus. *Am. J. Physiol.* **274:** C715–C723.

Smith, A.N., Wardle, C.J., and Harris, A. 1995. Characterization of DNase I hypersensitive sites in the 120kb 5′ to the CFTR gene. *Biochem. Biophys. Res. Commun.* **211:** 274–281.

Smith, A.N., Barth, M.L., McDowell, T.L., Moulin, D.S., Nuthall,

H.N., Hollingsworth, M.A., and Harris, A. 1996. A regulatory element in intron 1 of the cystic fibrosis transmembrane conductance regulator gene. *J. Biol. Chem.* **271:** 9947–9954.

Smith, S., Giriat, I., Schmitt, A., and de Lange, T. 1998. Tankyrase, a poly(ADP-ribose) polymerase at human telomeres. *Science* **282:** 1484–1487.

Sonnhammer, E.L. and Durbin, R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1–10.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids. Res.* **22:** 4673–4680.

Trapnell, B.C., Chu, C.S., Paakko, P.K., Banks, T.C., Yoshimura, K., Ferrans, V.J., Chernick, M.S., and Crystal, R.G. 1991. Expression of the cystic fibrosis transmembrane conductance regulator gene in the respiratory tract of normal individuals and individuals with cystic fibrosis. *Proc. Natl. Acad. Sci.* **88:** 6565–6569.

Trower, M.K., Orton, S.M., Purvis, I.J., Sanseau, P., Riley, J., Christodoulou, C., Burt, D., See, C.G., Elgar, G., Sherrington, R. et al. 1996. Conservation of synteny between the genome of the pufferfish (Fugu rubripes) and the region on human chromosome 14 (14q24.3) associated with familial Alzheimer disease (AD3 locus). *Proc. Natl. Acad. Sci.* **93:** 1366–1369.

Venkatesh, B., Si-Hoe, S.L., Murphy, D., and Brenner, S. 1997. Transgenic rats reveal functional conservation of regulatory controls between the *Fugu* isotocin and rat oxytocin genes. *Proc. Natl. Acad. Sci.* **94:** 12462–12466.

Venkatesh, B., Tay, B.H., Elgar, G., and Brenner, S. 1996. Isolation, characterization and evolution of nine pufferfish (*Fugu rubripes*) actin genes. *J. Mol. Biol.* **259:** 655–665.

Vuillaumier, S., Dixmeras, I., Messai, H., Lapoumeroulie, C., Lallemand, D., Gekas, J., Chehab, F.F., Perret, C., Elion, J., and Denamur, E. 1997. Cross-species characterization of the promoter region of the cystic fibrosis transmembrane conductance regulator gene reveals multiple levels of regulation. *Biochem. J.* **327:** 651–662.

White, N.L., Higgins, C.F., and Trezise, A.E.. 1998. Tissue-specific in vivo transcription start sites of the human and murine cystic fibrosis genes. *Hum. Mol. Genet.* **7:** 363–369.

Xiu-Bao, C., Tabcharanis, J.A., Yue-Xian, H., Jensen, T.J., Kartner, N., Noa, A., Hanrahan, J.W., and Riordan, J.R. 1993. Protein kinase A (PKA) still activates CFTR chloride channel after mutagenesis of all 10 PKA consensus phosphorylation sites. *J. Biol. Chem.* **268:** 11304–11311.

Yeo, G.S., Elgar, G., Sandford, R., and Brenner, S. 1997. Cloning and sequencing of complement component C9 and its linkage to DOC-2 in the pufferfish *Fugu rubripes*. *Gene* **200:** 203–211.

Yoshimura, K., Nakamura,H., Trapnell, B.C., Dalemans, W., Pavirani, A., Lecocq, J.P., and Crystal, R.G. 1991a. The cystic fibrosis gene has a "housekeeping"-type promoter and is expressed at low levels in cells of epithelial origin. *J. Biol. Chem.* **266:** 9140–9144.

Yoshimura, K., Nakamura, H., Trapnell, B.C., Chu, C.S., Dalemans, W., Pavirani, A., Lecocq, J.P., and Crystal, R.G. 1991b. Expression of the cystic fibrosis transmembrane conductance regulator gene in cells of non-epithelial origin. *Nucleic Acids Res.* **19:** 5417–5423.

Received April 6, 2000; accepted in revised form June 2, 2000.

Human Molecular Genetics, 2000, Vol. 9, No. 9 1415–1423

Disruption of two novel genes by a translocation co-segregating with schizophrenia

J. Kirsty Millar[1,+], Julie C. Wilson-Annan[1,§], Susan Anderson[1], Sheila Christie[1], Martin S. Taylor[1], Colin A. M. Semple[1], Rebecca S. Devon[1], David M. St Clair[2], Walter J. Muir[1,3], Douglas H. R. Blackwood[1,3] and David J. Porteous[1]

[1]Medical Genetics Section, Department of Medical Sciences, The University of Edinburgh, Molecular Medicine Centre and MRC Human Genetics Unit, both at Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK, [2]The Department of Mental Health, University Medical Building, The University of Aberdeen, Foresterhill, Aberdeen AB9 2ZD, UK and [3]Department of Psychiatry, Royal Edinburgh Hospital, Morningside Park, Edinburgh EH10 5HF, UK

Received 1 February 2000; Revised and Accepted 27 March 2000

DDBJ/EMBL/GenBank accession nos AF222980–AF222987

A balanced (1;11)(q42.1;q14.3) translocation segregates with schizophrenia and related psychiatric disorders in a large Scottish family (maximum LOD = 6.0). We hypothesize that the translocation is the causative event and that it directly disrupts gene function. We previously reported a dearth of genes in the breakpoint region of chromosome 11 and it is therefore unlikely that the expression of any genes on this chromosome has been affected by the translocation. By contrast, the corresponding region on chromosome 1 is gene dense and, not one, but two novel genes are directly disrupted by the translocation. These genes have been provisionally named Disrupted-In-Schizophrenia 1 and 2 (*DISC1* and *DISC2*). *DISC1* encodes a large protein with no significant sequence homology to other known proteins. It is predicted to consist of a globular N-terminal domain(s) and helical C-terminal domain which has the potential to form a coiled-coil by interaction with another, as yet, unidentified protein(s). Similar structures are thought to be present in a variety of unrelated proteins that are known to function in the nervous system. The putative structure of the protein encoded by *DISC1* is therefore compatible with a role in the nervous system. *DISC2* apparently specifies a non-coding RNA molecule that is antisense to *DISC1*, an arrangement that has been observed at other loci where it is thought that the antisense RNA is involved in regulating expression of the sense gene. Altogether, these observations indicate that *DISC1* and *DISC2* should be considered formal candidate genes for susceptibility to psychiatric illness.

INTRODUCTION

Schizophrenia is a serious and debilitating disease affecting ~1% of the population worldwide. There is compelling evidence from family, twin and adoption studies for a significant genetic basis to the disease (1). This has initiated searches directed at identification of the genetic component using methods such as linkage analysis, association studies of candidate genes and mapping of cytogenetic abnormalities in psychiatric patients, procedures which have been applied successfully to monogenic disorders. Psychiatric illnesses are more complex, however, and apparently result from the combined effects of multiple genes, with inheritance complicated by environmental factors (1). Consequently no genes involved in the aetiology of such illnesses have yet been definitively identified.

We are studying a large Scottish family (2) in which a balanced translocation segregates with major mental illness (maximum LOD = 6.0; D.H.R. Blackwood, A. Fordyce, M. Walker, E. Drysdale, J.K. Millar, D.M. St Clair, D.J. Porteus and W.J. Muir, manuscript in preparation), based on the hypothesis that the rearrangement has directly disrupted gene function, leading to psychosis. This family may be atypical due to the wide spectrum of disorders present (schizophrenia, schizoaffective disorder, bipolar affective disorder, unipolar affective disorder and adolescent conduct disorder). However, it is likely that identification of the genetic factor(s) involved in the aetiology of disease in these patients would facilitate discovery and understanding of the underlying genetic defects in unrelated psychotic individuals.

RESULTS

To clone the chromosome 1 breakpoint, a 2.5 kb *Eco*RI fragment from chromosome 11, containing the site of the breakpoint (3), was used to screen an *Eco*RI genomic library constructed from a translocation cell line. A 2.7 kb *Eco*RI fragment, corresponding to the translocation fragment from the derived 1 chromosome was obtained (Fig. 1A). This fragment was used to rescreen the same library, yielding a 7.3 kb clone, containing the site of the chromosome 1 breakpoint (Fig. 1A). The derived 11 fragment was obtained using the polymerase

+To whom correspondence should be addressed at: Medical Genetics Section, Department of Medical Sciences, The University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, Scotland, UK. Tel: +44 131 650 1000; Fax: +44 651 1059; Email: kirsty.millar@ed.ac.uk

§Present address: Molecular Genetics of Cancer Division, WEHI, Royal Melbourne Hospital, Victoria 3050, Australia

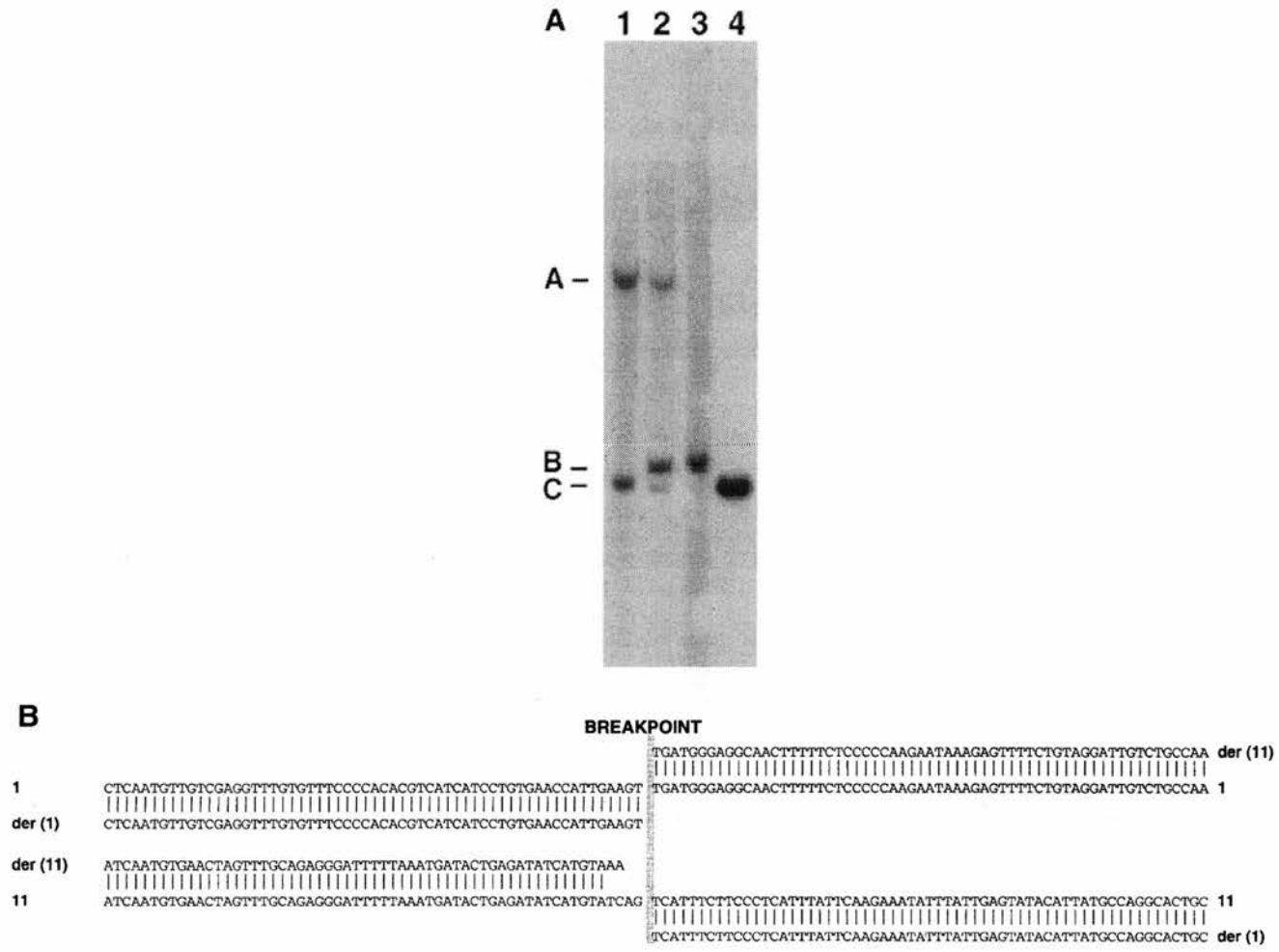


Figure 1. (**A**) Southern blot of *Eco*RI digested DNA probed with the 2.7 kb *Eco*RI derived 1 fragment. Lane 1, control human DNA; lane 2, genomic DNA from the translocation cell line MAFLI; lane 3, genomic DNA from the somatic cell hybrid MIS7.4, carrying the derived 1 translocation chromosome; lane 4: ICRF yeast artificial chromosome (YAC) y901 D0485 DNA. This YAC spans the chromosome 11 breakpoint of the translocation (28). Arrows indicate the positions of the 7.3 kb normal chromosome 1, 2.7 kb derived 1 and 2.5 kb normal chromosome 11 hybridization signals (A, B and C, respectively). (**B**) Alignment of sequence immediately flanking the breakpoints from the normal chromosome 1, derived 1, derived 11 and normal chromosome 11 [1, der (1), der (11) and 11, respectively]. The breakpoint sequence and minor rearrangement were confirmed by genomic sequence analysis of two other translocation carriers (data not shown). The position of the breakpoint was also confirmed by PCR and DNA sequence analysis on genomic DNA from MIS7.4 and MIS39, cell lines carrying the derived 1 and derived 11 chromosomes respectively (data not shown).

chain reaction (PCR). Alignment of breakpoint sequence from all four fragments (chromosome 11, chromosome 1, derived 11, derived 1, GenBank accession nos AF222984, AF222983, AF222986 and AF222985, respectively) (Fig. 1B) shows that the translocation resulted in replacement of TCAG with AA on the derived 11 chromosome, and that, consequently, no major rearrangement of sequence accompanied the translocation event.

Sequence analysis provided no evidence for the presence of any transcripts within the sequence surrounding the chromosome 11 breakpoint. However, the chromosome 1 breakpoint fragment sequence includes three expressed sequence tags (ESTs: AA249072, W04811 and D78808) and an exon from a messenger RNA, AB007926 (Fig. 2).

We have provisionally named the anonymous mRNA AB007926 *DISC1* (Disrupted-In-Schizophrenia 1), because analysis of the chromosome 1 breakpoint sequence indicates that the gene is directly disrupted by the translocation, which takes place within an intron (Fig. 2). Furthermore, this analysis indicates that the direction of transcription is proximal to distal.

Using a combination of cDNA library screening and RACE (rapid amplification of cDNA ends), we obtained 6913 nt of cDNA sequence transcribed from the *DISC1* gene (accession no. AF222980), which matched AB007926 (99.985%) over 6833 nt, and extended 14 nt further 5′. Four different 5′ RACE products, our longest cDNA clone and AB007926 all terminate within 14 nt, suggesting that this is the true 5′ end of the transcript. Furthermore, Nucleotide Identifiy X (NIX, http://menu.hgmp.mrc.ac.uk/

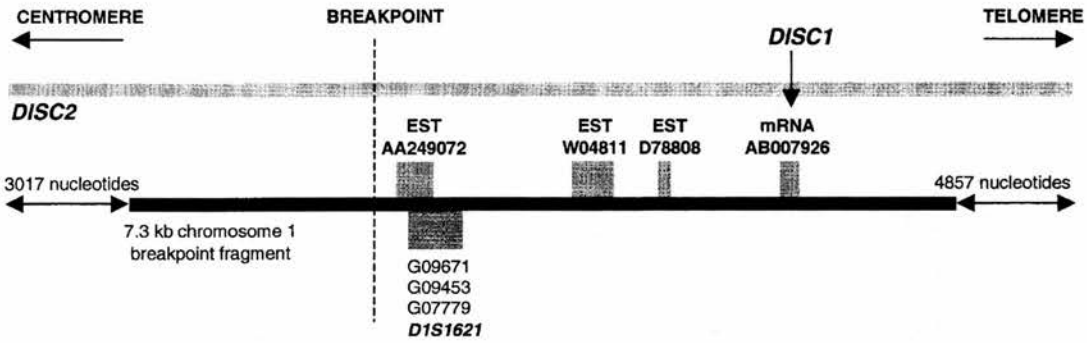


Figure 2. Map of the 7.3 kb *Eco*RI chromosome 1 breakpoint fragment. Breakpoint sequence was analysed using BLASTN (27) at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) and the suite of gene recognition and analysis programmes encompassed by NIX. Only 103 nt of the total 350 in EST D78808 are contained within the chromosome 1 breakpoint sequence. The remaining sequence is identical to several other ESTs (UniGene cluster IIs.31446, http://www.ncbi.nlm.nih.gov/UniGene/index.html), none of which contain any chromosome 1 breakpoint sequence or are even present on chromosome 1, as judged by a lack of hybridization to genomic DNA from the chromosome 1 human/mouse hybrid cell line A9(Neo-1)-4 (data not shown). The messenger RNA sequence in AB007926 consists of 6833 nt of a brain-expressed transcript (29) positioned on chromosome 1 by Gene Map '98 (http://www.ncbi.nlm.nih.gov/genemap/). 191 nt of this transcript are contained within the 7.3 kb of chromosome 1 breakpoint sequence. NIX identified one putative exon with consensus splice sites on the forward strand of the chromosome 1 breakpoint sequence. This exon contains 189 nt of the sequence match to mRNA AB007926. The sequence of a marker, *D1S1621*, is also contained within the chromosome 1 breakpoint fragment. The extent to which *DISC2* continues centromeric and telomeric from the chromosome 1 breakpoint fragment is indicated.

menu-bin/Nix/Nix.pl) analysis of 2.4 kb of genomic sequence encompassing the 5′ end of *DISC1* identified a 758 nt putative CpG island (70% GC) containing 120 nt from the 5′ end of the transcript, and two potential promoters 55 and 359 nt upstream. The first ATG in *DISC1* is located at position 54 and is not part of a strong translation initiation consensus (4). The predicted coding sequence, starting at the first ATG, consists of 2565 nt, with a stop codon at position 2616. The 3′ untranslated region (UTR) is 4294 nt extending to a poly(A) tail. A consensus polyadenylation signal is located at position 6892, 16 nt upstream of the poly(A) tail. The gene is tagged at the 3′ end by UniGene cluster Hs.26985.

Sixty-six nucleotides (2295–2360) are not contained within the putative coding sequence of AB007926. This deletion corresponds to a common alternative splicing event (Fig. 3, Table 1). *DISC1* is present as a major transcript of ~7.5 kb (Fig. 4A and B) in all adult tissues examined. The size discrepancy between the sequence we have obtained and the transcripts detectable on northern blots may be due to polyadenylation. *DISC1* was not detected on northern blots of fetal tissues (data not shown), although reverse transcription (RT)–PCR experiments indicate that fetal transcripts do exist (Table 1).

The open reading frame (ORF) in *DISC1* encodes a putative protein of 854 amino acids. Protein structure prediction programmes (http://dodo.cpmc.columbia.edu/predictprotein/) suggest that DISC1 can be divided into two distinct regions of secondary structure (Fig. 5A). The N-terminal region (amino acids 1–347) is predicted to consist of one or more globular domains. The C-terminal region is predicted to consist entirely of α-helix interspersed with several short loops, and contains regions with the potential to form coiled-coils, structures that arise when the helical stretches of separate proteins interact. Alternative splicing introduces an additional loop into the C-terminal domain.

BLAST 2.0 and FASTA 3 searches of the SWall database (SwissProt plus TREMBL) at the European Bioinformatics

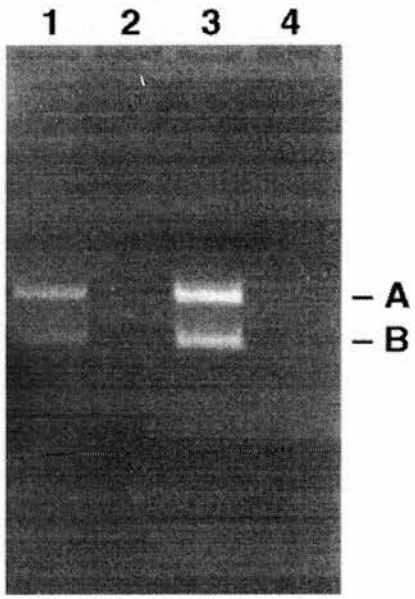


Figure 3. RT–PCR analysis of *DISC1* in human fetal brain using the distal primer pair (Table 1). Position 2295, at which the 66 bp deletion starts, corresponds approximately to a splice donor site within the cDNA (AGgtattg instead of the consensus AGgtaagt). Position 2360, where the deletion ends, is known to be a splice donor (J.K. Millar, S. Christie, D. Lawson, D. Hsiao-Wei Loh, B. Arveiler and D.J. Porteus, manuscript in preparation). A product of 270 bp indicates upstream donor site usage. A product of 336 bp indicates downstream donor site usage. Lane 1, 8.3 weeks; lane 2, 10.3 weeks. Arrows indicate the 336 and 270 bp alternatively spliced products (A and B, respectively).

Institute (http://www2.ebi.ac.uk/) reveal matches (~21% identity, 41% similarity across 500 amino acids) between the α-helical region of DISC1, particularly around the stretches of coiled-coil, and the known or predicted coiled-coil domains of

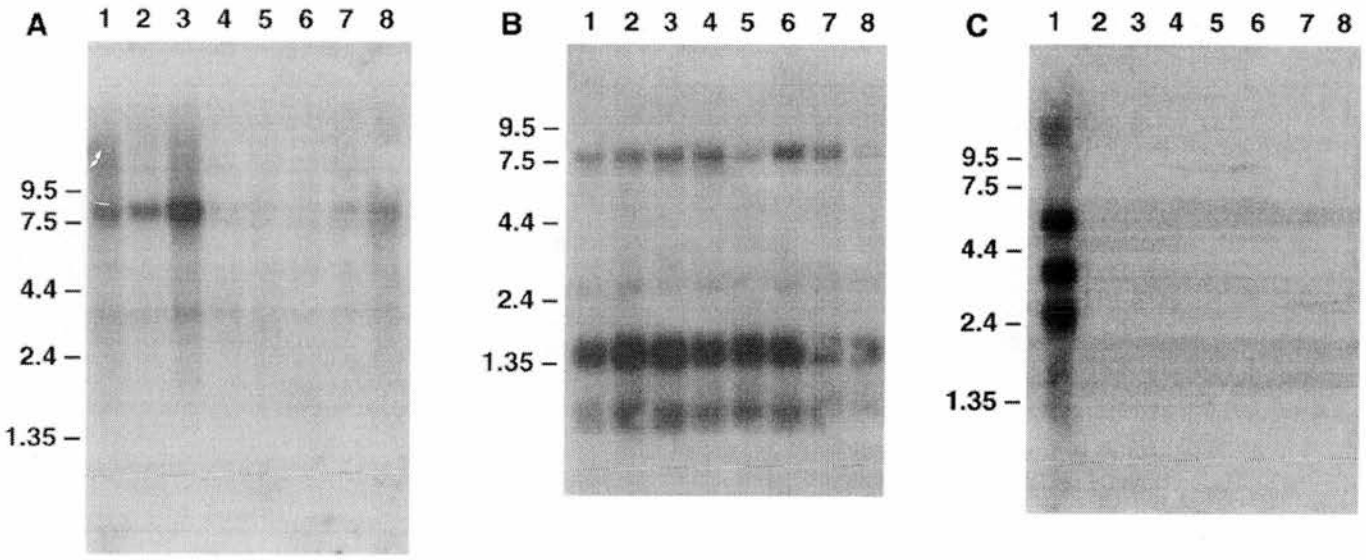


Figure 4. Clontech human multiple tissue northern blots probed with *DISC1* (nucleotides 958–1983) and *DISC2* (nucleotides 8488–9300). Each lane contains ~2 μg of poly(A)$^+$ RNA. Positions of size markers are indicated. (**A**) Adult human multiple tissue northern probed with *DISC1*. Lanes 1–8: heart, brain, placenta, lung, liver, skeletal muscle, kidney and pancreas. (**B**) Adult human brain multiple tissue northern probed with *DISC1*. Lanes 1–8: amygdala, caudate nucleus, corpus callosum, hippocampus, whole brain, substantia nigra, subthalamic nucleus and thalamus. An additional filter with RNA from cerebellum, cerebral cortex, medulla, spinal cord, occipital pole, frontal lobe, temporal lobe and putamen was probed and produced similar results (data not shown). (**C**) Adult human multiple tissue northern probed with *DISC2*. Lanes 1–8: heart, brain, placenta, lung, liver, skeletal muscle, kidney and pancreas. All four heart transcripts are apparently derived from *DISC2* because no other sites in the human genome to which *DISC2*-derived sequences hybridize have been detected (data not shown).

Table 1. RT–PCR analysis of *DISC1* and *DISC2*

Sample	Age (weeks)	*DISC1*		*DISC2*	
		proximal	distal	proximal	distal
Brain	8.3	+	+ (2)	+	+
	10.3	+	+ (2)	+	+
	13.3	+	+ (2)	+	+
Heart	8.8	+	+ (2)	+	+
	9.1	+	+ (2)	+	+
	9.3	+	+ (2)	+	+
Kidney	10.0	+	+ (2)	–	+
Spleen	14.8	+	+ (2)	+	–
Limb	10.3	+	+ (2)	–	–

Approximate ages of gestation are given in weeks. 2, two bands obtained using one primer pair indicating alternative splicing; +, transcript detected; –, transcript not detected.

several other proteins. These similarities are likely to result from biased sequence composition imposed by coiled-coil structure and therefore probably reflect structural resemblances between the proteins.

DISC2 (accession no. AF222981) was identified from the ESTs at the chromosome 1 breakpoint (Fig. 2) and extended by cDNA library screening, RACE and RT–PCR. This identified 15 178 nt of contiguous genomic sequence which is known to be transcribed. The transcript 3′ end is tagged by Unigene cluster Hs.96883. There are two consensus polyadenylation signals at positions 15 072 and 15 161, 107 and 18 nt upstream of the poly(A) tail.

The 5′ end of the transcript has not yet been located, but *DISC2* so far consists of a single large exon encompassing the translocation breakpoint and the 189 nt exon of *DISC1*. Sequence analysis indicates that the direction of transcription is distal to proximal. *DISC2* therefore overlaps with *DISC1*, but is transcribed in the opposite orientation. *DISC2* transcripts are most abundant in heart where species of >9.5 kb, and of ~6, 3 and 2.5 kb are present (Fig. 4C). RT–PCR indicates that *DISC2* is also transcribed in several fetal tissues (Table 1) although transcripts were not detectable on northern blots (data not shown).

No significant ORF has been identified within 15 178 nt of the *DISC2* transcript. The longest ORF deduced to be present encodes 57 amino acids, while the start codon of this ORF is not in a good context (4), suggesting that the sequence lacks any protein coding potential. Furthermore, a survey of the EMBL database (ftp://ftp.ebi.ac.uk/pub/databases/embl/release/) indicates that the longest known 3′ UTR in a human gene is 9280 nt (doublecortin), while the average length is 2131.83 nt ± 1368.09 (based on 861 full-length 3′ UTR sequences). Therefore, if *DISC2* is a protein-coding gene, it possesses a 3′ UTR in excess of 15 kb, substantially outside the size range of other known 3′ UTRs. In addition, *DISC2* possesses certain similarities to the 17 572 nt non-coding mRNA-like transcript *NTT* (5), as summarized (Fig. 5B). These observations suggest that *DISC2* is a non-coding structural RNA gene, although final confirmation awaits identification of the transcript 5′ end.

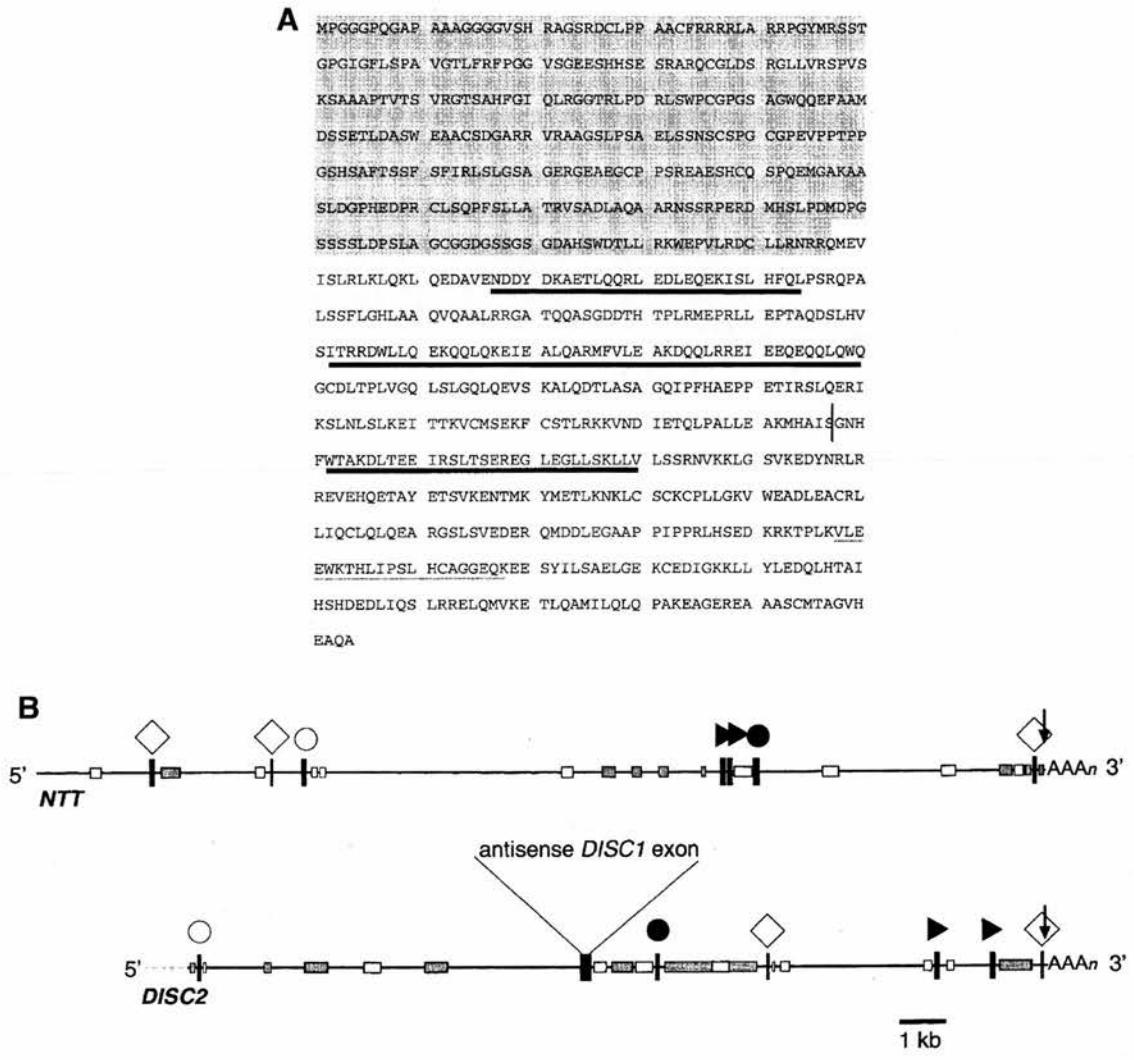


Figure 5. (**A**) Predicted protein sequence of DISC1 translated from the first ATG. Light underlining, alternatively spliced amino acids; bold underlining, stretches with coiled-coil forming potential; light shaded box, putative globular domain(s); vertical line, position of the translocation breakpoint. (**B**) Feature-based comparison of *DISC2* and *NTT* transcripts. Simple repeat and interspersed repeat features were annotated based on RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) analysis of each transcript. The direct repeat features were identified using Dotmatrix anaysis (30). Both transcripts and the annotated features are drawn to scale. The tetranucleotide repeat of *DISC2* is $(TGGA)_n$ and of *NTT* is $(TAGA)_n$. Although both transcripts possess comparable features, the linear organization of those elements is not highly similar. Open circle, tetranucleotide repeat; open diamond, AT- or GT-rich low complexity region; closed circle, polypurine tract; closed triangle, transcript-specific repeat element; vertical arrow, utilized poly(A) consensus; shaded box, LINE; open box, SINE; closed box, annotated feature.

DISCUSSION

We have cloned and sequenced the breakpoints of a (1;11)(q42.1;q14.3) translocation linked to schizophrenia and related psychosis (2), and identified two novel genes, *DISC1* and *DISC2*, both disrupted by the translocation. One or both of the disrupted alleles may be responsible for the psychiatric disorders suffered by carriers of the translocation. Furthermore, these genes may also be involved in the mental illness of patients unrelated to the family segregating the translocation. However, no independent evidence for a locus in this region of the genome has yet been presented, although recent reports of suggestive linkage to 1q32 and 1q32–41 in patients suffering from bipolar disorder and schizophrenia, respectively (6,7), are intriguing.

DISC1 protein matches are essentially restricted to structural similarities to myosins, structural proteins and proteins that are involved in motility and/or transport (particularly microtubule binding proteins). In the context of psychiatric illness, it is interesting to note that many of these proteins (dynactin, D-CLIP-190, citron, post-synaptic density proteins, FEZ1 and hyaluronan receptor, for example) are implicated in processes such as axon guidance, synaptogenesis, functioning of the synapse and intracellular transport along axons and dendrites (8–15). It is an intriguing possibility that the function of DISC1 is similar, suggesting a role in development of the nervous system and/or neuronal activity, and therefore adding further to the evidence pointing towards involvement in the aetiology of mental illness.

By analogy to many other examples of mammalian genes with endogenous antisense RNA transcripts (16–18), *DISC2* presents an attractive mechanism by which *DISC1* expression may be regulated. There is evidence that antisense RNAs affect expression of the sense gene (i) in the nucleus at the levels of transcription, RNA processing or export from the nucleus, or (ii) in the cytoplasm by influencing RNA stability or translation. Similarly, *DISC2* might act at one of these levels to influence expression of *DISC1*.

In translocation carriers, transcription of both *DISC1* and *DISC2* is predicted to occur from their endogenous promoters on the derived chromosomes, unless unidentified transcription signals of major effect are removed or inactivated by the rearrangement. This contention is supported by preliminary experiments which detect transcription of *DISC1* from the derived 1 chromosome (data not shown). It is unlikely that corresponding truncated transcripts lacking a 5′ end would be produced as well, because the breakpoint region of chromosome 11 is apparently transcriptionally inactive (3,19).

Several scenarios resulting from production of 3′ truncated *DISC1* and *DISC2* transcripts could be envisaged, but three are most obvious. First, truncated DISC1 protein may be produced. It would lack the C-terminal 257 amino acids, including one of the regions of most strongly predicted coiled-coil forming potential. This would be expected to reduce the overall coiled-coil forming potential and stability as a multimer, while retaining the unidentified function of the globular N-terminus. Production of a partially active protein could conceivably result in a dominant-negative effect. Second, truncated *DISC2* transcripts would retain complementarity to the normal full-length *DISC1* transcripts. However, *DISC2* truncation could affect events following interaction between truncated *DISC2* and normal *DISC1* transcripts. Therefore, if *DISC2* does regulate expression of *DISC1* it is possible that this mechanism would be negatively affected, resulting in dysregulation of both the non-translocated and translocated alleles, and a pleiotropy of dominant-negative effects. Third, irrespective of the putative regulatory role of *DISC2* on *DISC1*, the gene product may have independent (regulatory) functions affected by truncation.

We propose that alteration of *DISC1* and/or *DISC2* activity, by truncation and/or by abnormal regulation of expression, is causally linked to the psychiatric illness in translocation carriers. Dysregulation of the novel functions of *DISC1* and *DISC2* in the absence of a translocation event may play a more general role in susceptibility to psychiatric illness.

MATERIALS AND METHODS

Cell culture

The lymphoblastoid cell line MAFLI from an individual bearing the t(1;11)(q42.1;q14.3) translocation, somatic cell hybrids MIS7.4 and MIS39 bearing the derived 1 or derived 11 translocation chromosomes respectively, and their culture conditions, have been described previously (20). On the derived chromosome 1, DNA has been lost from 1q42.1-qter and replaced with chromosome 11 material from 11q14.3-qter. The derived 11 chromosome is the reciprocal translocated chromosome. The cell line A9(Neo-1)-4, a mouse A9 hybrid cell line carrying human chromosome 1, and its culture requirements, have been reported previously (21).

DNA preparation

Human and λ genomic DNA was prepared by standard methods (22). Plasmid and cosmid DNA was prepared using Qiagen (Crawley, UK) plasmid midi kits.

RNA extraction and cDNA synthesis

Human fetal tissues were obtained from the United Kingdom Medical Research Council Tissue Bank. Total RNA was extracted using RNazol B (Biogenesis, Poole, UK) according to the manufacturer's instructions. First strand cDNA synthesis was carried out on DNase I-treated RNA using the random hexamer primer from the SUPERSCRIPT Preamplification System (Gibco BRL, Paisley, UK) according to the manufacturer's instructions. One microlitre of the resulting cDNA was used in standard PCRs.

Genomic library construction and screening

Genomic DNA from the translocation cell line MAFLI was digested with *Eco*RI, ligated into *Eco*RI-digested and dephosphorylated λZAPII (Stratagene, Cambridge, UK), and packaged using Gigapack Gold II packaging extract (Stratagene) according to the manufacturer's instructions. Bacteriophage were plated using *Escherichia coli* XL1-Blue MRF' and the library of clones screened using standard methods. Excision of clones from the λ vector was carried out as advised by the manufacturer, releasing genomic fragments cloned into pBluescript SK(–). The library was screened using a 2.15 kb repeat-free *Hin*dIII–*Eco*RI sub-fragment of the 2.5 kb *Eco*RI fragment containing the chromosome 11 breakpoint (3), followed by the 2.7 kb derived 1 fragment.

cDNA library screening

5′-STRETCH PLUS cDNA libraries of 20–26 week fetal brain and 20–25 week fetal heart, constructed in bacteriophage λgt10 and gt11, respectively, were obtained from Clontech (Basingstoke, UK) and screened according to the manufacturer's instructions. Inserts were obtained from pure clones using two methods. First, cDNAs were amplified by PCR, turbocloned (23) and sequenced. In order to amplify cDNA inserts from λ vectors, a single plaque was picked into 25 µl of distilled water, 1–5 µl were then added to a PCR and the cDNA insert amplified using vector-based primers. λgt10-specific primers, agcaagttcagcctggttaagt and gggaccttctttatgagtatt (annealing at 68°C) and λgt11-specific primers gaaggcacatggctgaatatcgacggtttc and gacaccagaccaactggtaatggtagcgac (annealing at 56°C) were used to amplify inserts from the fetal brain and fetal heart cDNA libraries, respectively. Due to the probable introduction of sequence alterations during PCR, several subclones were sequenced. Alternatively, λ DNA was digested with *Eco*RI to release the cDNA insert which was then subcloned into *Eco*RI-digested pBluescript SK(–) (Stratagene).

DNA sequencing

In order to sequence the derived 11 PCR product, it was excised from a 0.8% low melting point agarose gel, dialysed in distilled water for 30 min, melted at 65°C and an appropriate quantity of melted gel slice added to the sequencing reaction.

Direct cosmid sequencing utilized 0.5–1 μg of cosmid DNA, with 60 ng of primer and ABI PRISM BigDye terminator cycle sequencing ready reaction kits (PE Applied Biosystems, Warrington, UK). Plasmid DNA sequencing reactions were performed using ABI PRISM dye terminator or dRhodamine terminator cycle sequencing ready reaction kits and 20 ng of primer. Products were separated on an ABI 373 or 377 DNA sequencer (PE Applied Biosystems), according to the manufacturer's instructions. Resulting sequence was analysed using the GCG package of sequence analysis software (Wisconsin package version 9.1, Genetics Computer Group, Madison, WI). Sequence contigs were constructed using the Phred, Phrap and Consed software, version 6.0 (24–26).

BLASTN (27) searches using chromosome 11 breakpoint sequence identified only sequence matches to two bacterial artificial chromosome (BAC) end clones (accession nos AQ748746 and AQ105798). Searches using chromosome 1 breakpoint sequence identified sequence matches to a triplet repeat *D1S1621* (accession nos G09671, G09453 and G07779), three BAC end clones (accession nos AQ112950, AQ078498 and B40542) as well as the mRNA and EST matches.

Sequencing of cosmid ICRFc112I0142Q6

This cosmid spans the chromosome 1 breakpoint (J.K. Millar, S. Christie, D. Lawson, D. Hsiao-Wei Loh, B. Arveiler and D.J. Porteus, manuscript in preparation). It was obtained by screening a chromosome 1 cosmid library provided by the Resource Centre/Primary Database (RZPD) of the German Human Genome Project at the Max-Planck-Institute for Molecular Genetics (http://www.rzpd.de). The probe used was the 2.7 kb derived 1 fragment. Cosmid DNA was treated with Plasmid-Safe ATP-dependent DNase (Epicentre Technologies, CAMBIO, Cambridge, UK) to remove contaminating *E.coli* DNA according to the manufacturer's instructions. Cosmid DNA was partially digested with *Sau*IIIA, and the resulting restriction fragments size-fractionated on a 0.8% agarose gel. Fragments of ~900 bp were excised from the gel and subcloned into pBluescript SK(–) (Stratagene). Subclones were picked randomly and sequenced using vector-based primers flanking the cloning site (caggaaacagctatgac and gtaaaacgacggccagt). Contig overlaps were established by designing primers from the ends of the contigs and subsequent direct cosmid sequencing. This process was repeated to generate the sequence of ICRFc112I0142Q6 as two ordered contigs (GenBank accession no. AF222987).

PCRs

Unless otherwise stated, PCR was carried out using Ampli*Taq* DNA polymerase (Perkin Elmer Biosystems, Foster City, CA). Each 50 μl reaction contained 1 U of enzyme, 300 ng of each primer, 200 mM of each dNTP, 1.5 mM $MgCl_2$, 50 mM KCl and 10 mM Tris–HCl pH 8.3. All reactions utilized 35 cycles with a denaturation step of 30 s at 94°C, an annealing step of 1 min at a temperature appropriate for the primers used, and a synthesis step at 72°C, based on the assumption that 1 min is required to synthesize 1 kb of DNA.

The 1.4 kb derived 11 breakpoint PCR product was amplified (annealing at 60°C) using one primer specific for chromosome 11 (ggctggatattgcccttgagccataatt) and one primer specific for chromosome 1 (agaacagaggagggacgatgatgac). MIS39 genomic DNA was used as template.

Analysis of *DISC1* expression by RT–PCR was performed using proximal primers ggaaggagcaggaggcagcccaggcgga and gcacgctgcaggtggtaagcaatc (152 bp product) with the Advantage-GC cDNA PCR kit (Clontech) and cycling conditions as recommended by the manufacturer. Distal primers ggaagcttgtcgattgcttatcc and agatcttcatcatgactgtggattgc designed from *DISC1* were also utilized for RT–PCR (270 and 336 bp products, annealing at 64°C). Analysis of *DISC2* expression by RT–PCR utilized proximal primers cccaagccttatccctcaggatcaa and atcaggcagaatatgccacagcgtg (250 bp product, annealing at 65°C) and distal primers gagacgacaagtcacagacttggag and gctctcaggcataagacactgtgac (486 bp product, annealing at 68°C). In the latter three cases, an initial hot start step was carried out.

Hybridization

Standard procedures were used for Southern blotting and hybridization (22). Double-stranded probes were labelled with [α-^{32}P]dCTP by random priming using High Prime (Boehringer Mannheim, Lewes, UK) and purified using Amersham Pharmacia Biotech (Little Chalfont, UK) NICK columns. The oligonucleotide probe was labelled with [γ-^{32}P]dATP using T4 polynucleotide kinase and hybridized to filters at 58°C.

Sequencing of *DISC1*

A 694 nt probe containing the 189 nt exon of AB007926 (contained within the 7.3 kb chromosome 1 breakpoint fragment) was prepared by PCR using the chromosome 1 breakpoint fragment as template and primers ccatttctggacggctaaagacc and gcaracactttggctaaggcggc (annealing at 58°C). This probe and subsequent rounds of cDNA library rescreening generated several overlapping clones from the fetal brain and fetal heart cDNA libraries spanning 6913 nt of the transcript. 5′ RACE products were obtained from *DISC1* using the Advantage-GC cDNA PCR kit (Clontech) and cycling conditions as recommended by the manufacturer, with human 20–25 week fetal brain Marathon-Ready cDNA (Clontech) as template (gene-specific primer 1, gagtcaaggccacactgtctggc; gene-specific primer 2, gcacgctgcaggtggtaagcaatc). Genomic sequence (2.4 kb) encompassing the 5′ end of *DISC1* was obtained as follows. An oligonucleotide (ggaaggagcaggaggcagcccaggcgga) designed from the 5′ end of *DISC1* was hybridized to a genomic *Pst*I fragment of ~2.4 kb. *Pst*I fragments of this size were isolated from a PAC, 135-G6, containing the 5′ end of *DISC1*, (J.K. Millar, S. Christie, D. Lawson, D. Hsiao-Wei Loh, B. Arveiler and D.J. Porteus, manuscript in preparation), subcloned into pBluescript SK(–) (Stratagene) using standard methods (22) and sequenced (GenBank accession no. AF222982).

Identification of *DISC2* transcribed sequence

Probes corresponding to ESTs AA249072 and W04811 were used to screen fetal heart and fetal brain cDNA libraries. A 555 nt probe for EST AA249072 was prepared using the chromosome 1 breakpoint fragment as template and primers gctgtcaattaagcagtaacagtgc and catctctgaaacagggatgtgtcc

(annealing at 58°C). The cDNA corresponding to EST W04811 was excised from modified pT7T3 (Pharmacia) by double-digestion with *Not*I and *Eco*RI. Further cDNA clones were identified using cDNAs corresponding to ESTs from Unigene cluster Hs.96883 to rescreen the fetal heart cDNA library (these ESTs were identified from sequencing of cosmid ICRFc112i0142Q6). *DISC2* was extended proximal and distal from the cDNA clones isolated by cDNA library screening using 5′ and 3′ RACE with human adult heart Marathon-Ready cDNA (Clontech) as template, and Expand LT (Boehringer Mannheim) with cycling conditions as recommended by the manufacturer. Gene-specific primers 1 and 2 were gcttgtgcttattctctttgggta, and accatcgtcactgtttctcctgct, gctttggcactttggttggctgta and tctttcacctcctcctcctcttctt, gccacccatgcccagctcactttta and gctctggcacatatttaaagaaagtatccc, tatttttccaggttctttccccag and tttcccttcttctctccacaacgt. Gaps were filled by speculative RT–PCR performed using primers designed along the length of the sequence of cosmid ICRFc112I0142Q6, and 10.0 or 13.6 week (gestational age) human fetal heart cDNA as template. Primer pairs (in order, proximal to distal along the cosmid sequence) were gtgggtaagggtattgtt and cacagagtttcagagttc, tgaggattgggcaggtgaaaggga and gctgactttacccactctgttcca, ttggaacagagtgggtaa and cagtgcccttggtaaaac, cttggggacatttttggcagg and tcaaatctggttcctatcagcc, tctcagaagacgctggttcagtgc and gtgaagtgacaatgatgagattc, cccttaagttggcttaacagctcag and tgtttcccacactctcaaccctag, atttgaggtgagttggcttttaggtg and tccttttctatcaccgtattgttct, aggtgcttgagttgtctgagttggat and tgaaaaactgctgcgtaaatctgagg, ggacagccctcagatttacgcagcag and acaaaatcttgctgcttggtattctc, ctaagtgagaataccaagcagcaaga and tcccttctctctcttttgcttcctc, actggtgagagaggaagcaaaaagag and tgtcccggctctccatttcctccctc, cttttattggcagggagagggaggaa and aaccccgatgacatgcaatttaccta, cagaagaaaatgccaatgtcaagtgt and caagccctaattcacctcgacagttt, aaactgtcgaggtgaattagggcttg and gccacatagcaccgcaacactcatct, ccagatgagtgttgcggtgctatgtg and aggcaaaaatggctgaagggaaataa, gattatttcccttcagccatttttgc and ttgtggaaaggatggatgggtgggtg. RT–PCR products were obtained from *DISC2* using Expand LT (Boehringer Mannheim) with cycling conditions as recommended by the manufacturer. The combination of cDNA library screening, RACE and RT–PCR identified 15 178 nt of contiguous genomic sequence which is transcribed. Within this genomic sequence, nucleotides 8213–12 161 and 14 181–14 780 correspond to cDNA clones, nucleotides 6245–8436 and 14 594–15 178 correspond to RACE products and nucleotides 1–6470 and 12 027–14 189 correspond to RT–PCR products. The 5′ end of the transcript, which is located distal to cosmid ICRFc112I0142Q6, has not yet been identified due to the lack of further genomic sequence for the design of primers and speculative RT–PCR.

Human 3′ UTR length analysis

3′ UTRs were determined for the KIAA subset of mRNA sequences, based on coding sequence annotations. 3′ ends of the determined UTRs were checked for the presence of a consensus poly(A) signal (AATAAA) within 400 nt of the presumed end. UTRs failing to meet this criterion were excluded from further analysis. Sequence coordinates for full-length 3′ UTRs were determined and poly(A) signals detected using in-house software.

Northern blot analysis

A probe corresponding to nucleotides 958–1983 of *DISC1* was obtained by excising one of the *DISC1* fetal brain library cDNA clones using *Eco*RI. The cDNA clone corresponding to EST W04811 was used to identify *DISC2* transcripts. Northern blots were obtained from Clontech.

ACKNOWLEDGEMENTS

We thank Kathy Evans for useful discussion and critical reading of this manuscript. This work was supported by Organon NV and the UK Medical Research Council.

REFERENCES

1. McGuffin, P., Owen, M.J. and Farmer, A.E. (1995) Genetic basis of schizophrenia. *Lancet*, **346**, 678–682.
2. St Clair, D., Blackwood, D., Muir, W., Carothers, A., Walker, M., Spowart, G., Gosden, C. and Evans, H.J. (1990) Association within a family of a balanced autosomal translocation with major mental-illness. *Lancet*, **336**, 13–16.
3. Millar, J.K., Brown, J., Maule, J.C., Shibasaki, Y., Christie, S., Lawson, D., Anderson, S., Wilson-Annan, J.C., Devon, R.S., St Clair, D.M. *et al.* (1998) A long-range restriction map across 3 Mb of the chromosome 11 breakpoint region of a translocation linked to schizophrenia: localisation of the breakpoint and the search for neighbouring genes. *Psychiatr. Genet.*, **8**, 175–181.
4. Kozak, M. (1984) Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.*, **12**, 857–872.
5. Liu, A.Y., Torchia, B.S., Migeon, B.R. and Siliciano, R.F. (1997) The human NTT gene: identification of a novel 17-kb noncoding nuclear RNA expressed in activated CD4+ T cells. *Genomics*, **39**, 171–184.
6. Detera-Wadleigh, S.D., Badner, J.A., Berrettini, W.H., Yoshikawa, T., Goldin, L.R., Turner, G., Rollins, D., Moses, T., Sanders, A.R., Karkera, J.D. *et al.* (1999) A high-density genome scan detects evidence for a bipolar-disorder susceptibility locus on 13q32 and other potential loci on 1q32 and 18p11.2. *Proc. Natl Acad. Sci. USA*, **96**, 5604–5609.
7. Hovatta, I., Varilo, T., Suvisaari, J., Terwilliger, J.D., Ollikainen, V., Arajarvi, R., Juvonen, H., Kokko-Sahin, M.L., Vaisanen, L., Mannila, H. *et al.* (1999) A genome-wide screen for schizophrenia genes in an isolated Finnish subpopulation, suggesting multiple susceptibility loci. *Am. J. Hum. Genet.*, **65**, 1114–1124.
8. Dillman, J.F., Dabney, L.P., Karki, S., Paschal, B.M., Holzbaur, E.L. and Pfister, K.K. (1996) Functional analysis of dynactin and cytoplasmic dynein slow axonal transport. *J. Neurosci.*, **16**, 6742–6752.
9. Abe, T.K., Tanaka, H., Iwanaga, T., Odani, S. and Kuwano R. (1997) The presence of the 50 kDa subunit of dynactin complex in the nerve growth cone. *Biochem. Biophys. Res. Commun.*, **233**, 295–299.
10. Waterman-Storer, C.M., Karki, S.B., Kuznetsov, S.A., Tabb, J.S., Weiss, D.G., Langford, G.M. and Holzbaur, E.L. (1997) The interaction between cytoplasmic dynein and dynactin is required for fast axonal transport. *Proc. Natl Acad. Sci. USA*, **94**, 12180–12185.
11. Ahmad, F.J., Echeverri, C.J., Vallee, R.B. and Baas, P.W. (1998) Cytoplasmic dynein and dynactin are required for the transport of microtubules into the axon. *J. Cell Biol.*, **140**, 391–401.
12. Lantz, V.A. and Miller, K.G. (1998) A class VI unconventional myosin is associated with a homologue of a microtubule-binding protein, cytoplasmic linker protein-170, in neurons and at the posterior pole of drosophila embryos. *J. Cell Biol.*, **140**, 897–910.
13. Zhang, W., Vazquez, L., Apperson, M. and Kennedy, M.B. (1999) Citron binds to PSD-95 at glutamatergic synapses on inhibitory neurons in the hippocampus. *J. Neurosci.*, **19**, 96–108.
14. Kuroda, S., Nakagawa, N., Tokunaga, C., Tatematsu, K. and Tanizawa K. (1999) Mammalian homologue of the *Caenorhabditis elegans* UNC-76 protein involved in axonal outgrowth is a protein kinase C zeta-interacting protein. *J. Cell Biol.*, **144**, 403–411.
15. Nagy, J.I., Price, M.L., Staines, W.A., Lynn, B.D. and Granholm, A.C. (1998) The hyaluronan receptor RHAMM in noradrenergic fibers contrib-

utes to axon growth capacity of locus coeruleus neurons in an intraocular transplant model. *Neuroscience*, **86**, 241–255.

16. Dolnick, B.J. (1997) Naturally occurring antisense RNA. *Pharmacol. Ther.*, **75**, 179–184.
17. Knee, R. and Murphy, P.R. (1997) Regulation of gene expression by natural antisense RNA transcripts. *Neurochem. Int.*, **31**, 379–392.
18. Constancia, M., Pickard, B., Kelsey, G. and Reik, W. (1998) Imprinting mechanisms. *Genome Res.*, **8**, 881–900.
19. Devon, R.S., Evans, K.L., Maule, J.C., Christie, S., Anderson, S., Brown, J., Shibasaki, Y., Porteous, D.J. and Brookes, A.J. (1997) Novel transcribed sequences neighbouring a translocation breakpoint associated with schizophrenia. *Am. J. Med. Genet.*, **74**, 82–90.
20. Fletcher, J.M., Evans, K., Baillie, D., Byrd, P., Hanratty, D., Leach, S., Julier, C., Gosden, J.R., Muir, W., Porteous, D.J. *et al.* (1993) Schizophrenia-associated chromosome 11q21 translocation–identification of flanking markers and development of chromosome 11q fragment hybrids as cloning and mapping resources. *Am. J. Hum. Genet.*, **52**, 478–490.
21. Minoru, K., Motoyuki, S., Hiroyuki, M., Hideto, Y. and Mitsuo, O. (1989) Construction of mouse A9 clones containing a single human chromosome tagged with neomycin-resitance gene via microcell fusion. *Jpn J. Cancer Res.*, **80**, 413–418.
22. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning. A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
23. Boyd, A.C. (1993) Turbo cloning—a fast, efficient method for cloning PCR products and other blunt-ended DNA fragments into plasmids. *Nucleic Acids Res.*, **21**, 817–821.
24. Ewing, B., Hillier, L., Wendl, M. and Green, P. (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
25. Ewing, B. and Green, P. (1998) Basecalling of automated sequencer traces using phred. II. Error probablities. *Genome Res.*, **8**, 186–194.
26. Gordon, D., Abajian, C. and Green, P. (1998) A graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.
27. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
28. Evans, K.L., Brown, J., Shibasaki, Y., Devon, R.S., He, L., Arveiler, B., Christie, S., Maule, J., Baillie, D., Slorach, E. *et al.* (1995) A contiguous clone map over 3 Mb on the long arm of chromosome 11 across a balanced translocation associated with schizophrenia. *Genomics*, **28**, 420–428.
29. Seki, N., Ohira, M., Nagase, T., Ishikawa, K., Miyajima, N., Nakajima, D., Nomura, N. and Ohara, O. (1997) Characterisation of cDNA clones in size-fractionated cDNA libraries from human brain. *DNA Res.*, **4**, 345–349.
30. Sonnhammer, E.L. and Durbin, R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1–GC10.

Mammalian Genome 11, 395–398 (2000).

Mammalian Genome
Incorporating Mouse Genome

Isolation and characterization of the mouse translin-associated protein X (*Trax*) gene

Rebecca S. Devon,* Martin S. Taylor, J. Kirsty Millar, David J. Porteous

MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK

Received: 15 October 1999 / Accepted: 20 December 1999

The translin-associated protein X (Trax) has been identified as a factor of unknown function that forms complexes with the protein translin (Aoki et al. 1997; Taira et al. 1998). Translin itself was originally identified as a protein that binds specifically to consensus sequences found at the breakpoints of many chromosomal translocations, potentially playing a role in the recognition of staggered DNA ends (Aoki et al. 1995). In addition, translin is transported to the nucleus after DNA damage and may thus function in DNA repair (Kasai et al. 1997).

In addition to binding DNA, translin is also capable of binding to highly conserved sequences in the 3′ untranslated regions (UTRs) of certain RNAs, thereby suppressing their *in vitro* translation. The mouse homolog of translin (also called testis-brain RNA binding protein, TB-RBP; Wu et al. 1997), together with an additional unidentified factor (possibly Trax), complex with a non-protein coding RNA called BC1 to form ribonucleoprotein particles (Muramatsu et al. 1998). These particles are translocated to neuronal dendrites in response to neuronal activity (Kobayashi et al. 1998) and are thought to play a role in local translational regulation, a process that is key to the maintenance and modulation of synaptic activity.

Here we describe the identification and preliminary characterization of *Trax,* the murine ortholog of the human TRAX gene. To derive the mouse *Trax* cDNA sequence, the mouse subset of dbEST was screened by BLASTN (Altschul et al. 1997), searching with the coding sequence of human TRAX (accession number X95073). A set of 11 highly homologous ESTs was identified, assembly of which into two non-overlapping contigs accounted for 86% of the expected coding region when aligned with human TRAX. dbEST was then searched iteratively with EST sequences derived from the opposite ends of the cDNA clones, and an apparently complete 3′ UTR sequence was also assembled.

In order to confirm the consensus mouse *Trax* sequence derived from the assembly of ESTs, cDNA clones constituting a minimal tiling path across the gene (clone numberes 445424, 805812, 476185, and 653425) were obtained from the UK-HGMP Resource Centre and their inserts fully sequenced. This resulted in a complete sequence, minus one gap of 178 bp, which was covered only by poor quality EST sequence from a cDNA clone that was not publicly available. This gap was therefore filled by performing RT-PCR across the gap and sequencing of the product. The complete sequence is 2390 bp long up to the point of poly(A) addition. The sequence has been submitted to the Genbank database with the accession number AF187040 and to the Mouse Genome Database (MGD) with the accession number J:58681.

By analogy with human TRAX, the coding sequence is presumed to begin at base 201 within exon 1; this is supported by the presence of a near-perfect consensus Kozak sequence neighboring the initiator methionine codon, the absence of upstream ATG codons, and the presence of an in-frame stop codon in the 5′ UTR. The open reading frame is 870 nucleotides long and encodes a 290-amino acid protein, which is the same length as that encoded by the human gene. The 3′ UTR extends for 1302 bp before a consensus poly(A) addition signal (AATAAA) is encountered, and a poly(A) tail is added 13 bases later.

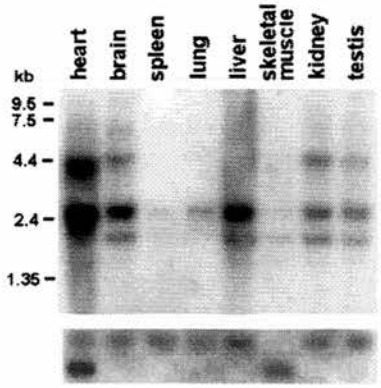


Fig. 1. Expression pattern of mouse *Trax*. Northern blot (purchased from Clontech) containing 2 μg poly(A)+ RNA from various adult mouse tissues, hybridized with a 249-bp probe derived from the mouse *Trax* coding sequence (upper panel). The scale in kilobases is shown to the left. The lower panel shows the same blot hybridized with a control β-actin derived probe. The expected 2.0-kb band is seen in all tissues except skeletal muscle. The smaller band visible for heart and skeletal muscle is due to cross-reaction of the probe with other forms of actin.

A probe derived from the coding sequence of mouse *Trax* was used to probe Northern blots to determine transcript size. Three major bands were detected, of approximately 2 kb, 2.5 kb, and 4.5 kb, with considerable variation in the intensity of each between tissues (Fig. 1). The human Trax gene also gives rise to multiple transcripts (Millar et al. in preparation), although reported originally in the literature as a single transcript of 2.7 kb (Aoki et al. 1997). The 2390-bp mouse *Trax* cDNA sequence described here is presumed to correspond to the 2.5-kb band. The longer transcript may be due to the presence of an extended 3′ UTR, as indicated by positive RT-PCR results for immediate downstream genomic sequence (data not shown). By analogy, the shortest, 2-kb isoform may correspond to a truncated 3′ UTR; there are two potential non-consensus poly(A) addition sites (AAGAAA at base 1895, AATAAG at base 1923) within the 3′ UTR that could give rise to a transcript of this size. It is interesting to note that while the 3′ UTR of mouse *Trax* is well represented among ESTs, there is no evidence for ESTs derived from the shortest or longest isoform. This is most surprising since, from the Northern blot analysis, the isoforms appear to be transcribed in approximately equal ratios in

Correspondence to: R.S. Devon; *e-mail:* Rebecca.Devon@ed.ac.uk

**Present address:* Medical Genetics Section, Molecular Medicine Centre, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH42XU, UK

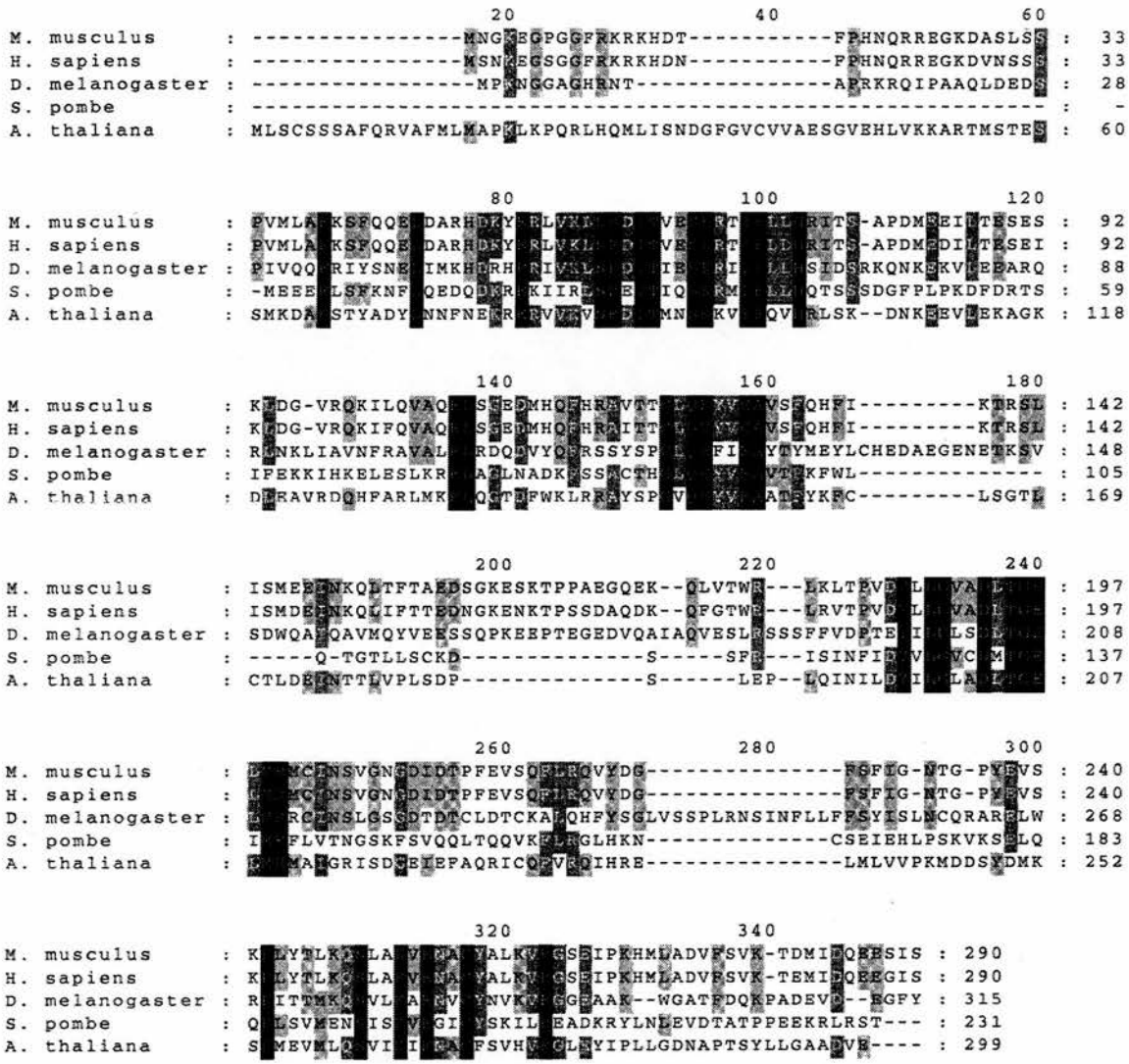


Fig. 2. Evolutionary conservation of TRAX. Alignment (ClustalW) of TRAX amino acid sequence in mouse, human, *D. melanogaster, A. thaliana,* and *S. pombe.* Residues colored black are identical in all five species, those in dark grey are identical between four species, and those in light grey are identical among three.

kidney and testis, and the 4.5-kb isoform is highly expressed in heart. A possible explanation is that secondary structure in the RNA molecule derived from the shortest and longest isoform hinders access of the reverse transcriptase enzyme required to make cDNA.

Expression of mouse *Trax* was assessed by hybridization to Northern blots, RT-PCR, and analysis of the tissues of origin of corresponding ESTs. Northern blots demonstrated expression in all adult tissues tested (Fig. 1) with higher levels of expression in heart and liver and lower levels in lung and skeletal muscle. Thirty-five cycles of RT-PCR on a wide range of adult tissues and developmental stages demonstrated approximately equal expression levels in each (data not shown). Examination of the tissue of origin of mouse *Trax* ESTs confirmed these findings and provided evidence for expression in further tissues. This wide-ranging expression profile suggests that mouse *Trax* is expressed ubiquitously throughout development of the embryo and continues to be expressed in all adult tissues.

The 870-nucleotide open reading frame of mouse *Trax* shares 86% identity with human TRAX at the nucleotide level, and the 290-amino acid-encoded protein exhibits 90% identity and 94% similarity (with Blosum62 matrix) to human TRAX at the amino acid level (Fig. 2). Comparison with more distantly related species shows 35%, 34%, and 30% amino acid identity between mouse and *D. melanogaster, A. thaliana,* and *S. pombe* respectively, with 11% of mouse residues being absolutely conserved between all of these species (M.S. Taylor et al., manuscript in preparation). The majority of absolutely conserved residues are clustered into four regions of ungapped alignment (Fig. 2).

In addition to a high degree of cross-species conservation within the coding sequence, evolutionary conservation can also be observed within the 3′ UTR. Above the background level of conservation, two nucleotide elements stand out as being particularly highly conserved (nucleotides 1889–1930 and 2063–2189 in mouse), exhibiting 86% and 80% nucleotide identity respectively between human and mouse. Neither of the elements matches any recognized sequence element, and there is no obvious palindromicity, repetitiveness, or stem-loop forming potential. Their function is unknown, although they may represent control elements concerned with mRNA stability, translational efficiency, or localization on the cytoskeleton.

It should also be noted that mouse *Trax* may be considered a paralog of translin, the protein to which it is known to bind, since it is 24% identical and 41% similar to mouse translin at the amino acid level. This level of homology is largely consistent for mouse *Trax* compared with translin orthologs from other species, sug-

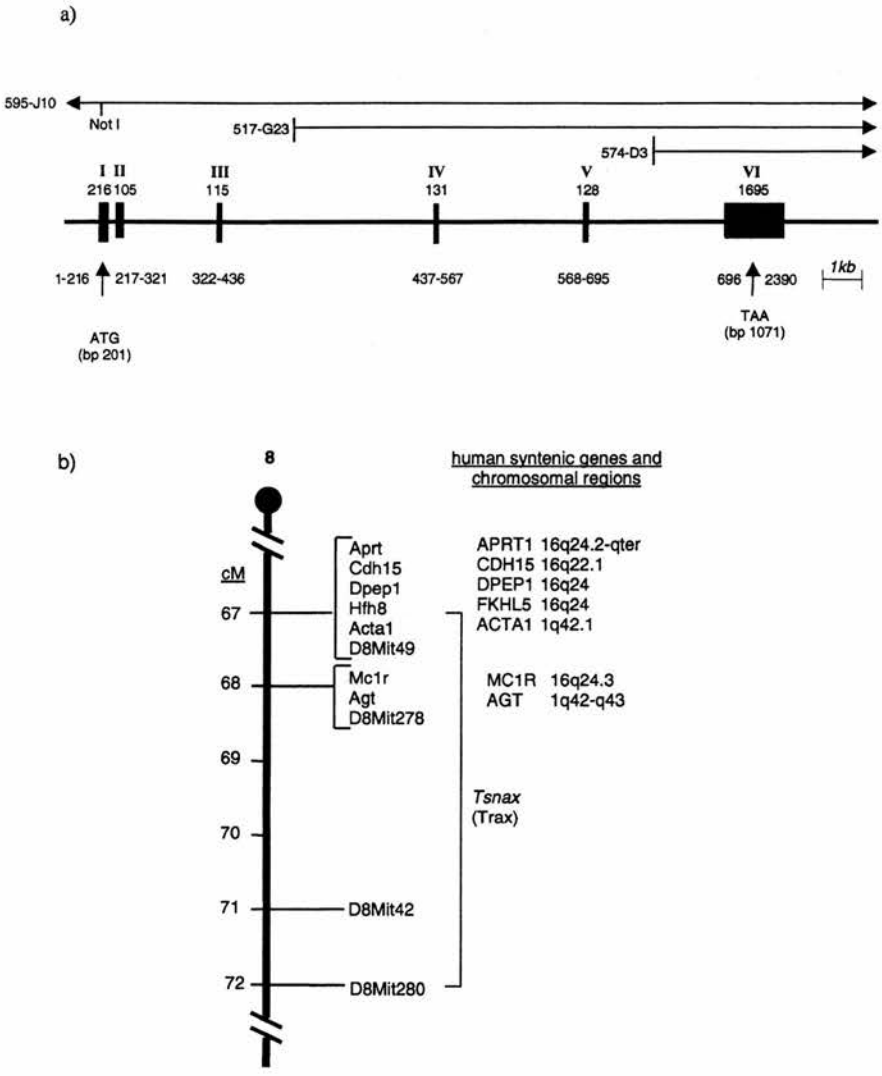


Fig. 3. a) Genomic organization of mouse *Trax*. Exons are shown as black boxes with the exon number and size in base pairs above and position in the cDNA sequence below the box. PACs are depicted as thin black lines. The start codon ATG is within exon 1 at base 201, and the stop codon TAA is within exon 6 at base 1071. Intron sizes are drawn to scale and range from 297 bp to approximately 5.5 kb. Purified PAC DNA was used for the determination of genomic structure of the mouse *Trax* gene by a variety of methods: a) direct sequencing on PAC DNA with cDNA-derived primers; b) PCR amplification across introns on PAC DNA with primers from adjacent exons and sequencing of products; and c) vector-hexamer PCR adapted from the method of Herring et al. (1998). To construct a physical map of the region, we isolated PAC end clones by modified vector-hexamer PCR and mapped them onto Southern blots of restriction-digested PAC DNA, and the products of *Not*I restriction-digested PAC DNA were resolved by pulsed field gel electrophoresis. **b)** Chromosomal localization of mouse *Trax*. Schematic of mouse Chr 8 showing the map location of *Trax*. Distance down the chromosome in cM is shown to the left. The position of previously mapped genes and markers plus conservation of synteny with human genes is shown to the right. A polymorphic dinucleotide (CA) repeat and complex $GT_{(2-5)}$ repeat containing clone was serendipitously isolated from a *Sau*3AI subclone library of PAC 574-D3. A single pair of primers was designed that flanked both these repeats (5′ AAA GGA AAG TTT GTA CCT GCC 3′ and 5′ CTT AGC GGA CCT CAA AGG CTC 3′), generating a 467-bp product in genomic DNA, and this PCR assay was used to determine the chromosomal location of *Trax* by linkage analysis on the CXB RI strains obtained from The Jackson Laboratory, Bar Harbor, Maine, USA.

gesting that the interaction between TRAX and translin is not in the form of an extensive dimerizing interface that would be expected to show co-ordinated substitutions between the two proteins.

In order to investigate the genomic structure of mouse *Trax*, the RPCI21 gridded mouse PAC library (obtained from the UK-HGMP Resource Centre) was screened with a probe derived from the genomic region of human TRAX. Three positive PACs were obtained (clone numbers 595-J10, 517-G23, and 574-D3) and used to determine the gene's intron and exon sizes and splice site sequences. The mouse *Trax* gene was found to span a distance of approximately 20 kb in genomic DNA and to comprise six exons of size ≥216, 105, 115, 131, 128, and ≥1695 bases respectively (Fig. 3a). All splice sites conformed to the consensus dinucleotides AG at the 3′ end of each intron and GT at the 5′ end (data not shown).

Southern blot analysis was used to demonstrate that mouse *Trax* is present as a single copy in the mouse genome (data not shown). The chromosomal localization of mouse *Trax* was then established by PCR and typing of a neighboring polymorphic repeat on the CXB recombinant inbred (RI) strains. The strain distribution pattern (SDP) of this polymorphism was identical to that at two markers within the genome, *D8Mit278* and *D8Mit42*, which enabled the boundaries of the map location of *Trax* to be defined as between *D8Mit49* (located 67cM from the centromere of the chromosome) at the proximal boundary and *D8Mit280* (located at 72cM) at the distal boundary, a distance of 5cM (Fig. 3b). The map location was confirmed by FISH analysis of PAC 595-J10 DNA on mouse metaphase chromosomes (data not shown) and has been submitted to the Mouse Genome Database with the locus symbol *Tsnax*.

This region in the mouse genome lies at a boundary of conservation of synteny between human Chromosomes (Chrs) 1 and 16. The neighboring mouse genes alpha actin 1 (*Acta1*) and angiotensinogen (*Agt*) both have orthologs on human Chr 1q42, in a 5-cM region between *D1S439* and *D1S459*, although other genes in this region are syntenic with human Chr 16q22-q24. The map location of human TRAX is not yet available on the Gene Map of the Human Genome. However we have recently mapped it to human Chr 1q42 (J.K. Millar et al. manuscript in preparation). This is consistent with the observed localization of mouse *Trax* and compounds our sequence-based evidence that the sequence described here represents the true ortholog of the human gene.

Acknowledgments. We thank Fiona Kilanowski for performing the FISH analysis, Brendan Innes for providing DNA from the CXB RI strains, and Julie Wilson-Annan and Kathy Evans for providing critical comments on the manuscript. This work was supported by the Medical Research Council.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402

Aoki K, Suzuki K, Sugano T, Tasaka T, Nakahara K et al. (1995) A novel

gene, *translin,* encodes a recombination hotspot binding protein associated with chromosomal translocations. Nat Genet 10, 167–173

Aoki K, Ishida R, Kasai M (1997) Isolation and characterisation of a cDNA encoding a translin-like protein, TRAX. FEBS Lett 401, 109–112

Herring CD, Chevillard C, Johnston Sl, Wettstein PJ, Riblet R (1998) Vector-hexamer PCR isolation of all insert ends from a YAC contig of the mouse Igh locus. Genome Res 8, 673–681

Kasai M, Matsuzaki T, Katayanagi K, Omori A, Maziarz RT et al. (1997) The translin ring specifically recognizes DNA ends at recombination hotspots in the human genome. J Biol Chem 272, 11402–11407

Kobayashi S, Takashima A, Anzai K (1998) The dendritic translocation of translin protein in the form of BC1 RNA protein particles in developing rat hippocampal neurons in primary culture. Biochem Biophys Res Commun 253, 448–453

Muramatsu T, Ohmae A, Anzai K (1998) BC1 RNA protein particles in mouse brain contain two y-, h-element-binding proteins, translin and a 37kDa protein. Biochem Biophys Res Commun 247, 7–11

Taira E, Finkenstadt PM, Baraban JM (1998) Identification of Translin and Trax as components of the GS1 strand-specific DNA binding complex enriched in brain. J Neurochem 71, 471–477

Wu X-Q, Gu W, Meng XH, Hecht N (1997) The RNA-binding protein, TB-RBP, is the mouse homologue of translin, a recombination protein associated with chromosomal translocations. Proc Natl Acad Sci USA 94, 5640–5645